

Inhoudstafel

Ten Geleide	3
Hoofdstuk 1 Wat is statistiek?	9
1.1. Statistiek als ‘gevaarlijk’ hulpmiddel	11
1.2. Statistiek = ?	13
RESPONSEN	19
Hoofdstuk 2 Het statistisch programma R	23
2.1. R omgeving	25
2.2. R installeren	25
2.3. Installatie van pakketten (<i>packages</i>)	29
2.4. Werken met pakketten ‘ <i>packages</i> ’	30
2.5. Conventies in R	32
Hoofdstuk 3 Data en de datamatrix	35
3.1. Wat is data en wat zijn variabelen?	37
3.2. Het meetniveau van variabelen	38
3.3. De datamatrix	47
RESPONSEN	50
Hoofdstuk 4 Databeheer in R	55
4.1. Soorten data	57
4.2. Het aanmaken van data	59
4.3. Een databestand aanmaken	61
4.4. Werken in en met een dataframe	63
4.5. Bestaande datasets inlezen	68
4.6. Bestaande functies inlezen	70
RESPONSEN	71
Gehanteerde functies	75
Hoofdstuk 5 De frequentieverdeling van een variabele	77
5.1. Absolute en relatieve frequenties	79
5.2. Frequentietabel	81
5.3. Cumulatieve frequenties	82
5.4. Histogram	88
5.5. Grafische voorstellingen van categorische variabelen	92
RESPONSEN	106
Gehanteerde functies	115

Uitgeverij Academia Press
Prudens Van Duyseplein 8
9000 Gent
België

www.academiapress.be

Uitgeverij Academia Press maakt deel uit van Lannoo Uitgeverij, de boeken- en multimediasdivisie van Uitgeverij Lannoo nv.

ISBN 978 94 014 5092 8
D/2017/45/709
NUR 916

Sven De Maeyer, Tine van Daal & Ellen Vandervieren
Univariate statistiek voor de menswetenschappen – Een Open Leerpakket in R
Gent, Academia Press, 2017, 260 p.

Vormgeving cover: Keppie en Keppie
Vormgeving binnenwerk: Punctilio

© Sven De Maeyer, Tine van Daal, Ellen Vandervieren & Uitgeverij Lannoo nv, Tielt

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotokopie, microfilm of op welke wijze ook, zonder voorafgaande schriftelijke toestemming van de uitgever.

Hoofdstuk 6 Parameters van ligging en spreiding	117
6.1. Parameters van ligging	119
6.2. Parameters van spreiding	131
6.3. Grafische weergave van ligging en spreiding: de boxplot	145
RESPONSEN	149
Gehanteerde functies	159
Hoofdstuk 7 Parameters van vorm	161
7.1. Scheefheid	163
7.2. Platheid (Kurtosis)	167
RESPONSEN	170
GEHANTEERDE FUNCTIES	175
Hoofdstuk 8 De (standaard-)normaalverdeling	177
8.1. De normaalverdeling	179
8.2. Z-scores	187
RESPONSEN	196
Gehanteerde functies	200
Hoofdstuk 9 Steekproeftheorie	201
9.1. Wat is een populatie?	203
9.2. Steekproeven	205
9.3. De ene steek is de andere niet	207
9.4. Fouten in steekproeven	213
RESPONSEN	221
Gehanteerde functies	228
Hoofdstuk 10 Inferenties over de verdeling van variabelen in de populatie	229
10.1. Betrouwbaarheidsintervallen rond het gemiddelde	231
10.2. Betrouwbaarheidsintervallen rond de variantie	240
10.3. Betrouwbaarheidsintervallen voor de kengetallen van vorm	246
10.4. Betrouwbaarheidsintervallen voor relatieve frequenties	248
RESPONSEN	250
Gehanteerde functies	260

Ten Geleide

Alvorens dit boek te gebruiken is het belangrijk om een aantal afspraken te maken en toe te lichten hoe dit boek is opgevat en opgebouwd. We doen dit onder de vorm van enkele topics over dit boek.

1. Waarover hebben we het in dit boek?

In wetenschappelijk onderzoek heeft de statistiek een belangrijke plaats ingenomen. Dit geldt ook voor de menswetenschappen. Statistiek is een krachtig instrument dat vaak te pas en te onpas wordt ingezet om wetenschappelijke kennis te onderbouwen uit zeer diverse menswetenschappelijke disciplines: bv. psychologie, sociologie, politieke wetenschappen, pedagogie, taal- en letterkunde,... Dit boek geeft een inleiding in de meest frequent gebruikte concepten en technieken uit de univariate statistiek voor menswetenschappen. Het heeft daarbij tot belangrijkste doel het inzicht van de lezer te verhogen. We starten daarbij van nul en bouwen het statistisch inzicht uit tot het infereren naar een populatie toe van univariate statistieken.

Statistiek maakt vaak gebruik van wiskunde en is eigenlijk een wiskundige toepassing. Dit heeft voor veel mensen tot gevolg dat ze het gebruik van statistiek ontwijken aangezien ze niet goed zijn "met cijfers". In dit boek proberen we de verschillende concepten en technieken op een niet-wiskundige wijze toe te lichten en te drenken in realistische voorbeelden. In die zin heeft dit boek voornamelijk de toepassing van statistiek voor ogen eerder dan de droge theorie achter de statistiek. Dit neemt niet weg dat we die theorie daar waar nodig niet links laten liggen.

De toepassing van de statistiek staat dus centraal in dit boek. Om statistiek toe te passen bestaan verschillende al dan niet commerciële software-pakketten (vb. R, SPSS (PASW), SAS, STATA,...). In dit boek maken we gebruik van R. Enkele redenen liggen aan de grondslag van deze keuze. Een eerste reden is het vrij beschikbaar zijn van dit softwarepakket. Het is open-source software die bovendien op verschillende besturingssystemen werkt (Windows, MacOs, Linux). Dit maakt dat dit in letterlijke zin het meest toegankelijke softwarepakket is. Bovendien is het in R mogelijk om zeer diverse analyse technieken toe te passen. De gemeenschap van statistici werkt bijna elke nieuwe techniek uit naar een toepassing in R. Daardoor kunnen de meest recente en gevorderde analysetechnieken vaak

ook toegepast worden in R, wat in andere softwarepakketten niet altijd het geval is.

2. Voor wie is het boek bedoeld?

Dit boek is bedoeld voor zowel studenten als onderzoekers al dan niet uit de academische wereld. Om het boek te kunnen hanteren is geen statistische voorkennis nodig. Het enige wat je onder de knie moet hebben zijn simpele rekenkundige bewerkingen: optellen, aftrekken, vermenigvuldigen, delen, machten en wortels. Voor onderzoekers kan het boek een goede manier zijn om hun kennis op te frissen of bij te benen en om te leren werken met het veelzijdige pakket R.

3. Hoe is het boek ingedeeld?

Het boek kan je grofweg in vijf delen opsplitsen. Het eerste deel maakt je wegwijs in wat we verstaan onder statistiek en laat je ook kennismaken met het softwarepakket dat we zullen hanteren om de statistische analyses uit te voeren (Hoofdstukken 1 en 2). In het tweede deel lichten we toe wat variabelen zijn, de verschillende soorten variabelen en hoe je deze variabelen binnen het pakket R kan beheren (Hoofdstukken 3 en 4). Deel drie reikt je de technieken aan om zicht te krijgen op hoe één bepaald kenmerk verdeeld is binnen de groep van eenheden die je onderzoekt (Hoofdstukken 5, 6, en 7). In een vierde deel van het boek staan we stil bij hoe we vanuit de beschrijvende analyses op steekproefgegevens meer te weten kunnen komen van de verdeling van een variabele in de hele populatie (Hoofdstukken 8, 9 en 10).

Doorheen het OLP wordt gewerkt met datasets en een bestand dat specifiek geschreven functies voor dit OLP bevat. Deze bestanden zijn te downloaden op de open leeromgeving (Moodle) van Academia Press: <http://moodle.academiapress.be>.

Op diezelfde plaats kan je tevens aanvullend oefenmateriaal terugvinden.

4. Wat verstaan we onder een “Open Leerpakket”?

De filosofie achter dit boek is dat het je in staat zou moeten stellen om zelfstandig te leren. Tijdens het lezen zullen we vaak beroep doen op jou als lezer om eerst na te denken vooraleer verder te gaan. In die zin heeft

het boek als doel je actief aan het werk te zetten als leerder. Zo wijkt het af van een klassiek boek doordat we met jou als lezer in dialoog gaan.

Al naargelang je voorkennis kunnen delen overbodig of onnodig expliciet overkomen. Door dit boek op een zelfstandige wijze door te nemen kan je zelf bepalen welke delen voor jou belangrijk, interessant, uitdagend,... zijn. Je kiest zelf wat je grondig doorneemt en waar je doorheen wandelt.

5. Wat is de betekenis van de verschillende gebruikte symbolen en lettertypes?

In dit boek maken we gebruik van verschillende symbolen die we telkens bij alinea's zetten. Deze symbolen geven aan wat de lezer mag verwachten in de bijhorende alinea. Hieronder de gebruikte symbolen en hun betekenis.



Een stukje **informatie**, dat je best zeer grondig doorneemt vooraleer verder te gaan.



Een **opdracht om zelf uit te voeren**; achteraan elk hoofdstuk kan je de responslaag voor de opdracht terugvinden.



Voor de analyses maken we gebruik van het softwarepakket R. Alinea's met dit symbool geven aan hoe de behandelde **inhoud toegepast kan worden in het softwarepakket R**.

Naast deze verschillende symbolen maken we ook gebruik van typografie. R wordt aangestuurd door commando's die we kunnen intypen. Deze commando's zullen we altijd weergeven in het ‘monospaced’ lettertype *courier new*.

R gebruikt “>” om een nieuwe lijn aan te geven waarin je het commando kan ingeven. Wanneer we dus een commando aangeven in dit boek zullen we dit als volgt doen:

```
> mean(Var1)
```

De pijl naar rechts “>” hoeft u echter niet meer als commando te typen in R zelf. In het boek geven we dus weer hoe het eruit zou moeten zien in R, inclusief de “>”.

Wanneer R een nieuwe lijn aangeeft, gebeurt dit door het plus-teken “+”. Wanneer de commando's te lang worden voor één regel zullen we ook

hiervan gebruik maken analoog met R. Om de leesbaarheid te verhogen zullen er soms extra spaties worden toegevoegd in commando's. Deze hoeft u niet mee over te nemen, ze hebben geen invloed op de werking. Hoofdletters en kleine letters hebben echter wel een invloed.

Andere namen, documenten en menu's staan in *cursief*.

Menu's en keuzes worden als volgt beschreven: *File > Save as*, waarmee we aangeven "kies 'Save as' uit het keuzemenu 'File'."

6. Is het belangrijk dat je weet wat er gebeurt bij een statistische analyse?

De volgende "parabel" (met dank aan prof. dr. H. van den Bergh) heeft als belangrijkste boodschap: gebruik enkel statistische technieken indien je weet wat je aan het doen bent. Deze boodschap lijkt simpel. Desalniettemin wagen onderzoekers zich vaak aan het uitvoeren van statistische technieken zonder ze echt te doorgronden. De filosofie van dit boek is hoofdzakelijk je dat inzicht mee te geven. We hopen daarin te slagen....

Een groepje van drie statistici trekt samen naar een statistisch congres in Munchen. Ze verzamelen op het perron van Antwerpen Centraal. Een van hen herkent een collega-onderzoeker en stapt op hem af.

"Wat doe jij hier?" vraagt de statisticus.

"We verzamelen hier met drie onderzoekers die samen naar een congres gaan over kwalitatieve onderzoekstechnieken in Munchen", antwoordt de collega kwalitatief onderzoeker.

"Ah, en hoeveel treinkaartjes hebben jullie gekocht?", vraagt de statisticus.

"Drie uiteraard!", antwoordt z'n collega.

"Hmm, mooi. Wij hebben er één bekocht."

"Hè?! Hoe doen jullie dat dan straks op de trein?", vraagt de andere.

"Oh, eenvoudig. Wij als statistici hebben zo onze methodes.", zegt de statisticus.

Beide groepjes stappen op en nemen dicht bij elkaar plaats. Ergens halfweg is het moment aangebroken. De statistici zien in de verte de controleur aankomen. Als de bliksem verdwijnen ze met z'n drieën en nemen plaats in het toilet. De controleur komt langs, klopt op de deur en vraagt: "Uw kaartje graag". Waarop één van de statistici hun enige

kaartje onder de deur schuift, de controleur het een knipje geeft en vervolgens doorgaat. De kwalitatieve onderzoeker heeft dit geobserveerd en denkt daaruit inzicht te hebben ontwikkeld in hoe dit werkt. Een week later zien ze elkaar opnieuw op het perron in Munchen.

"Hey, we hebben nu ook maar één kaartje gekocht!", zegt de kwalitatieve onderzoeker niet zonder enige trots.

"Oh, mooi." zegt de statisticus. "Wij hebben er geen gekocht".

Helemaal uit z'n lood geslagen vraagt de kwalitatieve onderzoeker: "Hoe gaan jullie dat doen?"

"Tja, wij hebben zo onze methodes.", zegt de statisticus.

Ze stappen samen op en nemen dicht bij elkaar plaats. Ergens rond Luxemburg zien ze de controleur in de verte opdagen. Als de bliksem verdwijnen de kwalitatieve onderzoekers samen in de toilet. Waarop de statistici opstaan, en één van hen op de deur van de toilet klopt en zegt: "Ihre Ticket bitte". De kwalitatieve onderzoekers schuiven hun ticketje onder de deur, waarop de statisticus het aanneemt en met z'n collega's in het andere toilet kruipst.

HOOFDSTUK 1

Wat is statistiek?

DOELSTELLINGEN:

Na dit hoofdstuk:

- ben je bewust van de gevaren die schuilen in het gebruik van statistiek;
- ken je de verschillende situaties waarbij statistiek kan helpen.



Het woord statistiek is voor vele mensen een gekend woord. Naast het feit dat het de haren doet oprijzen van de gemiddelde persoon, is het zo dat mensen met dit woord vaak naar verschillende zaken verwijzen. In dit hoofdstuk staan we stil bij wat wij precies verstaan onder statistiek, zetten we kort uiteen waarom een goede kennis van statistiek onontbeerlijk is en geven we aan wat de mogelijkheden zijn van statistiek.

1.1. Statistiek als ‘gevaarlijk’ hulpmiddel

1.1.1 We starten dit boek met drie voorbeelden van het gebruik van statistiek.



Voorbeeld 1

Op de website van de vrt werd om de andere dag een poll georganiseerd. Hiermee wou de vrt de mening van haar publiek opmeten ten aanzien van diverse topics. Heel vaak zie je dat duizenden mensen deelnemen aan dergelijke polls. De website gaf telkens wel de aantallen weer, maar zegt niets over de waarde van die poll.

Hoe ziet zo’n peiling eruit? De onderstaande illustratie toont een voorbeeld van zo’n poll:

Moeten de universiteiten en hogescholen overheids geld krijgen per geslaagde student in plaats van per ingeschreven student, zoals nu het geval is?

aantal stemmen: 3023

435 (14%)

ja, want zo zullen ze meer moeite doen om ook de zwakkere studenten te helpen slagen

2588 (86%)

nee, want zo zullen ze de studenten te makkelijk laten slagen en zullen ze af willen van de zwakkere studenten

Figuur 1.1: Voorbeeld van een Poll op de website van VRT



1.1.2 a) Kan je volgens jou uit de bovenstaande peiling afleiden dat 86% van de Vlamingen vindt dat de universiteiten en hogescholen geen overheids geld moeten krijgen per geslaagde student?

b) In juni 2005 heeft de VRT een peiling georganiseerd op haar website over de houding van de Vlamingen ten aanzien van de Europese Grondwet. De poll werd nooit getoond.

Waarom doen ze dat volgens jou? Waarom worden sommige resultaten getoond, terwijl anderen worden geweerd?

1.1.3 Voorbeeld 2



President Bush Jr. van de Verenigde Staten had allerlei wilde plannen met de belastingen voor de Amerikanen vooraleer hij verkozen werd. In het voorjaar van 2003 zei dhr. Bush letterlijk: "*Under this plan, 92 million Americans receive a tax cut of \$1083*". Want volgens zijn berekeningen zou het gemiddelde belastingsvoordeel \$1083 bedragen.

1.1.4



Kan je volgens jou uit het feit dat het gemiddelde belastingsvoordeel \$1083 bedraagt, afleiden dat de gemiddelde Amerikaan een belastingsvoordeel van \$1083 zal hebben onder het nieuwe plan? Zo neen, wat zou je uit deze woorden wel kunnen afleiden?

1.1.5 Voorbeeld 3



Een zekere Steven Levitt (econoom) en John Donohue (Jurist) stelden in 2001 dat de legalisering van abortus in de VS vanaf 1973 een invloed had op de mate van criminaliteit, met enige vertraging (nl. 20 jaar)¹.

Op basis van de analyse van misdaadcijfers van 1985 tot 1997, samen met de analyse van abortusgegevens vanaf 1973, bekeken ze patronen van afname in criminaliteit. Deze afname stemt overeen met de periode waarin de kinderen die in de jaren van legalisering zijn geboren, in hun late adolescentie komen. De staten die abortus het eerst legaliseerden, zijn de staten waarin misdaad ook het eerst afnam. De staten met de hoogste mate van abortus zijn ook de staten met de grootste afname in misdaadcijfers.

Het rapport kan je vinden op de CD-Rom in het mapje "Achtergrondliteratuur".

1. Donohue III, J. en Levitt, S. (2001). "The Impact of Legalized Abortion on Crime." *Quarterly Journal of Economics*, 2001, 116(2), pp. 379-420.

1.1.6 Ben je akkoord met de conclusie van de auteurs dat het legaliseren van abortus leidt tot minder misdaad?



1.2. Statistiek = ?

1.2.1



Het woord *statistiek* heeft dezelfde etymologische stam als staat. Als specialiteit werd het ontwikkeld in de periode waarin de moderne natie-staat allerlei instrumenten ontwikkelde om vat te krijgen op de sociale omgeving. In eerste instantie had statistiek te maken met de gegevens die de staat nodig dacht te hebben, gebruikte, om zijn beleid op af te stemmen.

Wat is een staat? Een staat bestaat uit een grondgebied, waarover macht wordt uitgeoefend door allerlei instellingen, en die zijn beslissingen in laatste instantie kan afdwingen door te berusten op geweld. Een staat heeft dus een geweldsmonopolie (of streeft dit alleszins na) en tracht de natuurlijke en sociale omgeving te controleren. Een staat tracht de macht aanvaardbaar te laten zijn door beslissingen te nemen die een zekere waarde hebben voor de onderdanen (of groepen van onderdanen). Eén van de belangrijke instrumenten hierin is het beschikken over informatie en het beheren van sleutelsectoren via informatie.

Moderne staten wilden bijvoorbeeld weten hoeveel mensen er woonden op hun grondgebied. Daarnaast wilden ze weten hoeveel mensen er geboren werden, stierven, verhuisden, waar ze werkten, wat ze deden voor werk, hoeveel geld er in omloop was,... Vanuit de nood naar dit soort informatie is de statistiek ontstaan. Volkstellingen waren de eerste belangrijke statistische instrumenten van een staat.

Daarnaast werden statistieken en statistische technieken ontwikkeld door mensen en organisaties die er winst uit konden slagen. Actuarissen, verzekeringen, banken,... ontwikkelden technieken om geldstromen in kaart te brengen, winstmaximaliserende strategieën te bedenken, levenskansen te berekenen, risico's van verzekerde cargo's in te schatten, ...

1.2.2



Als je zelf het woord statistiek(en) hoort, waaraan denk je dan spontaan?

1.2.3

 In welke van de volgende situaties kan statistiek een rol spelen en waarom wel of niet?

- Om te voorspellen dat een appel naar beneden valt eens hij losraakt van de boom;
- Om te weten dat kogels door een vitaal orgaan dodelijke verwondingen kunnen veroorzaken;
- Om te voorspellen dat een kind met een hoog IQ hoge cijfers voor rekenen behaalt op school;
- Om te weten dat een val van de bovenste verdieping van de Eifeltoren een dodelijke afloop kent.

1.2.4

 Statistiek kan drie verschillende functies hebben:

1. Beschrijven
2. Verklaren
3. Voorspellen

1.2.5

Beschrijven



In eerste instantie verzamelden staten eenvoudige gegevens over de bevolking om te weten wie die bevolking is. Statistiek dient dan om een vereenvoudiging te geven van een complexe realiteit, zoals de leeftijdssamenstelling van een populatie.

In 1846 werd, op initiatief van Adolphe Quetelet, de eerste volkstelling in België georganiseerd. Hoewel deze tellingen in de eerste plaats een administratieve doelstelling hadden, waren ze van meet af aan ook een belangrijke bron van informatie over de demografische en socio-economische kenmerken van de Belgische bevolking. Anderhalve eeuw later waren tellingen als dusdanig overbodig geworden toen het Rijksregister van de natuurlijke personen de voor de hand liggende bron werd om het bevolkingscijfer te bepalen. Het beschrijft het aantal burgers en dit kan worden gebruikt om bijvoorbeeld de kiesdelers te berekenen. We vergeten deze eerste eenvoudige functie maar al te vaak.

Binnen het domein van de menswetenschappen worden beschrijvingen van een groep mensen gebruikt om meer zicht te krijgen op de eigenschappen van die groep.

1.2.6

 Bedenk zelf een situatie bij een concreet (onderzoeks)probleem in het domein van onderwijs en/of opleidingen waarbij we de beschrijvende functie van statistiek zouden kunnen gebruiken.

1.2.7

Verklaren



Met statistieken kan je een **statistisch model** bouwen. Dit is een grove vereenvoudiging van de realiteit, waarin je beschrijft hoe situaties in gemiddelde termen / in probabilistische termen, werken. Statistiek kan met andere woorden worden ingezet om een bepaald fenomeen dat we vaststellen in de werkelijkheid te verklaren.

Een voorbeeld maakt dit duidelijker.

Er bestaat een verband tussen het roken van tabak en longkanker. Onder de conditie dat alle andere voorwaarden dezelfde zijn (*ceteris paribus*), heeft iemand die 20 sigaretten per dag rookt 20 keer meer kans om longkanker te krijgen dan een niet-roker. Een meer precies model (meer gedetailleerd model) gaat accurater zijn en bijgevolg ook realiteitsgetrouw zijn, tegen de prijs van complexiteit. Men kan bijvoorbeeld in het model rekening houden met sekse, leeftijd, passief roken, ander risicogedrag,... Dit model gaat een schatting geven van het aantal rokers dat longkanker zal ontwikkelen in vergelijking tot niet-rokers, en kan gebruikt worden om schattingen te maken over de last voor de sociale zekerheid,...

In dit voorbeeld proberen we het fenomeen "longkanker krijgen" te verklaren. Welke factoren verklaren dat mensen longkancers vormen?

1.2.8

 Bedenk zelf een situatie bij een concreet (onderzoeks)probleem in het domein van onderwijs en/of opleidingen waarbij we de verklarende functie van statistiek zouden kunnen gebruiken.

1.2.9

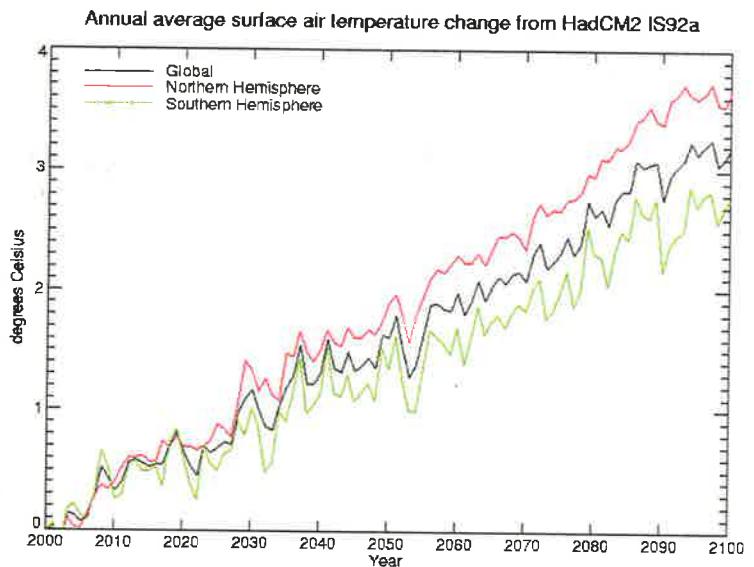
Voorspellen



Een derde mogelijke functie van statistiek is het formuleren van voorspellingen over wat kan gebeuren.

Een voorbeeld van een voorspelling die je kan maken met behulp van statistieken is de gemiddelde temperatuur over een heel jaar in het jaar 2100. De aarde warmt op en de gemiddelde temperaturen zullen tegen 2100

met 3,4° Celsius zijn toegenomen. Statistici hebben deze prognose uitgewerkt en samengevat in de onderstaande grafiek (Fig. 1.2).



Figuur 1.2: Verwachte opwarming van de aarde voor de periode 2000-2100

Een ander voorbeeld zijn de levensverwachtingen. De levensverwachting van een man die 36 jaar oud is en die in het Brussels Hoofdstedelijk Gewest woont, bedraagt op dit moment nog een extra 40,85 jaren. Dat kunnen we aflezen uit de onderstaande tabel.

Dit is de statistisch beste gok naar hoeveel jaar een 36-jarige nog voor de boeg heeft. Moest diezelfde persoon in Vlaanderen leven, zou zijn levensverwachting er nog beter uitzien (44,72 jaren).

Leeftijd (x)	Sterftekans (Qx)	Overlevingskans (Px)	Aantal overlevenden op 1.000.000 geboren (Lx)	Aantal sterfgevallen van de ene leeftijd tot de volgende (Dx)	Verwachte levensduur (Ex)
0	0,004530	0,995470	1.000.000	4.530	75,47
1	0,000985	0,999015	995.470	981	74,81
2	0,000243	0,999757	994.489	241	73,88
3	0,000204	0,999796	994.248	203	72,90
4	0,000313	0,999687	994.045	312	71,91
5	0,000162	0,999838	993.734	161	70,94
6	0,000110	0,999890	993.573	109	69,95
7	0,000112	0,999888	993.464	112	68,96
8	0,000227	0,999773	993.352	226	67,96
9	0,000114	0,999886	993.126	113	66,98
10	0,000345	0,999655	993.013	342	65,99
11	0,000231	0,999769	992.671	230	65,01
12	0,000117	0,999883	992.442	116	64,02
13	0,000118	0,999882	992.325	117	63,03
14	0,000181	0,999819	992.208	179	62,04
15	0,000183	0,999817	992.028	182	61,05
16	0,000249	0,999751	991.847	247	60,06
17	0,000251	0,999749	991.600	249	59,08
18	0,000697	0,999303	991.351	691	58,09
19	0,000685	0,999315	990.660	679	57,13
20	0,000660	0,999340	989.981	653	56,17
21	0,000624	0,999376	989.328	618	55,21
22	0,000712	0,999288	988.711	704	54,24
23	0,000960	0,999040	988.007	948	53,28
24	0,000706	0,999294	987.059	697	52,33
25	0,000896	0,999104	986.362	884	51,37
26	0,000667	0,999333	985.478	658	50,41
27	0,000760	0,999240	984.820	748	49,45
28	0,000970	0,999030	984.072	954	48,48
29	0,001008	0,998992	983.118	991	47,53
30	0,000908	0,999092	982.127	891	46,58
31	0,000938	0,999062	981.235	920	45,62
32	0,001055	0,998945	980.315	1.034	44,66
33	0,000727	0,999273	979.281	712	43,71
34	0,001426	0,998574	978.569	1.396	42,74
35	0,001246	0,998754	977.173	1.218	41,80
36	0,000974	0,999026	975.955	950	40,85
37	0,001545	0,998455	975.005	1.506	39,89

Figuur 1.3: Sterftetafels 2001-2003 Brussels hoofdstedelijk gewest (mannen)

Nog een ander voorbeeld. De kans om als voetganger door een auto te worden omver gereden bedroeg in 2002 in de VS 1 op 47 273, terwijl de kans om te sterven in een vliegtuigongeval 1 op 40 951 bedroeg in hetzelfde jaar (wat eigenlijk evenveel kans is om te sterven aan de impact van een asteroïde van een diameter van ten minste 2 km en met een energetische impact van meer dan 1 miljoen megaton TNT). De kans om de lotto te winnen is nog kleiner.

Statistiek kan echter **niet exact voorspellen**. Het zijn geen **definitieve voorspellingen**, maar wel voorspellingen in termen van **kansen**.

Als ik een 36-jarige ben uit het Brussels Hoofdstedelijk Gewest dan is mijn levensverwachting een gemiddelde levensduur. Het is de gemiddelde levensduur van vele mannen van mijn leeftijd. Dit impliceert niet dat ik exact nog 40 jaar zal leven. Ik kan morgen doodvallen, maar ik kan ook nog een extra 60 jaren leven.

Responsen

Respons 1.1.2

- a) Eigen aan zo'n polls is dat je niet met een representatief staal van de Vlaamse bevolking zit.

Je kan enkel uitspraken doen over de groep mensen die de website van de vrt bezocht heeft en die bereid is om aan zo'n peiling deel te nemen. Dit is een zeer specifieke groep van Vlamingen. Bijgevolg kunnen we de cijfers die we in zo'n poll vaststellen niet zo maar veralgemenen als de mening van de Vlamingen als groep.

Dit vormt meteen het verschil met wetenschappelijk onderzoek. Binnen wetenschappelijk onderzoek zullen we ernaar streven om een duidelijke groep van mensen af te bakenen waarover we een uitspraak willen doen. Dit zullen we verder de populatie noemen. Vervolgens zullen we leren hoe we te werk gaan om een zo representatief mogelijk staal van leden van die populatie te betrekken in het onderzoek door op één of andere manier (hoe zien we later) een representatieve steekproef te trekken.

- b) Eigen aan zo'n poll's is dat ze vaak niet meteen de resultaten opleveren die verwacht waren en bijgevolg soms aan nieuwswaarde verliezen. Of de resultaten blijken naderhand politiek te gevoelig zijn om te publiceren. Bijgevolg wordt er selectieve censuur uitgeoefend.

Dit is ook sterk in tegenstelling tot wat we verwachten van wetenschappelijk onderzoek. Indien uit het toepassen van een statistische techniek blijkt dat we een andere resultaat dan verwacht bekomen, dan wordt bij wetenschappelijk onderzoek verwacht dat ook daar verslag van gemaakt wordt. Dit wordt dan vaak aangevuld met alternatieve verklaringen voor de resultaten.

Respons 1.1.4

Dit is een verkeerde conclusie. Het Brookings Institution Tax Policy Centre rekende uit dat 80% van de Amerikanen niet zoveel belastingsaftrek zou krijgen. De middelste 20% van de belastingbetalers zou een gemiddelde belastingsvermindering van \$256 krijgen, terwijl bijna de helft van alle belastingbetalers minder dan \$100 belastingsvermindering zou krijgen.

Dit komt omdat er een beperkte groep rijke Amerikanen is die een zeer groot belastingsvoordeel zou hebben onder het nieuwe regime. Het belastingsvoordeel voor

deze groep Amerikanen trekt het gemiddelde belastingsvoordeel enorm omhoog. Er zijn dus verschillende manieren om gemiddelden te berekenen. De president gebruikt het gewone rekenkundige gemiddelde en vindt een waarde van \$1083. Later zullen we zien dat de mediaan van belastingsvermindering in dit geval een veel eerlijker beeld naar de bevolking zou hebben gegeven. De mediaan ligt echter bij \$100 en is dus beduidend minder spectaculair om campagne mee te voeren.

Uit dit verhaal leren we dat statistiek vaak misbruikt wordt voor verschillende doelstellingen. Statistiek wordt vaak gebruikt om bepaalde statements meer zeggingskracht te geven of krachtiger te maken. Maar zoals in het voorbeeld van de president geeft statistiek maar een zeer partiële samenvatting van de werkelijkheid.

Dit is te vergelijken met het beeld van Godfried Bomans over de statistiek, nl.
"Een statisticus waadde vol vertrouwen door een rivier die gemiddeld één meter diep was. ... Hij verdronk."

Respons 1.1.6

De gepresenteerde studie is een observatiestudie op basis van secundaire gegevens. Een gevolg daarvan is dat andere kenmerken van de staten niet kunnen worden opgenomen in de analyse. Het is best mogelijk dat het vastgestelde verband toe te schrijven is aan andere kenmerken die de staten waar abortus gelegaliseerd is gemeen hebben. We kunnen het echter niet nagaan. Zijn het net niet die meer librale staten die ook een veel grondiger beleid gevoerd hebben om de misdaad te doen dalen? Wat met bijvoorbeeld de hogere mate van gevangennname? Deze cijfers zijn namelijk ook toegenomen. Is er een invloed van programma's zoals de "war on drugs", ...?

Uit dit voorbeeld leren we dat de interpretatie van de resultaten van statistische analyse altijd met de nodige voorzichtigheid moet gebeuren. Er zijn altijd alternatieve verklaringen die niet kunnen worden gecontroleerd bij statistische analyses.

Respons 1.2.2

Vandaag kan het woord "statistiek" drie verschillende ladingen dekken:

1. Statistieken over inkomens, aantal leerkrachten lager onderwijs, aantal leerlingen lager onderwijs, opgetekende temperaturen per dag, het aantal doelpunten gescoord door een bepaalde speler of door een welbepaalde voetbalploeg, ...

Hier staat het woord statistiek gelijk aan eenvoudige lijsten van cijfers. Het gaat om enkelvoudige brokken uniforme informatie. Dit zullen we doorgaans "gegevens" noemen. We zullen het ook vaak aanduiden met de Latijnse benaming "data".

2. Statistiek kan ook duiden op de resultaten van berekeningen op deze elementaire data. Het gemiddelde inkomen van mannen en vrouwen in 2010, het gemiddelde aantal leerlingen per leerkracht in het lager onderwijs in het schooljaar 2009-2010, de spreiding van het aantal doelpunten over het seizoen, ...
3. Ten slotte wordt het woord statistiek gebruikt om te duiden op de tak van de wetenschap die ons helpt om data te analyseren, en om conclusies eruit te trekken. Het vak statistiek gaat hierover.

Respons 1.2.3

Enkel in situatie c) kunnen we statistiek en waarschijnlijkheidsrekenen gebruiken. Dit zijn instrumenten die pas hun waarde ten volle krijgen als je te maken hebt met onzekerere situaties. Als we volledig zeker zijn en alle elementen volledig voorspelbaar zijn, dan heb je geen statistieken nodig. We gaan bijvoorbeeld niet spreken van de kans dat een appel van een boom naar beneden valt eens hij losraakt, die kans is namelijk één, hij valt altijd. De zwaartekracht is een voorspelbare grootheid, toch zeker op de aarde. Je hebt ook geen statistieken nodig om te weten of kogels door een vitaal orgaan dodelijke verwondingen kunnen teweegbrengen of om te weten dat een val van de bovenste verdiepingen van de Eifeltoren een dodelijke afloop kent.

Enkel als we te maken hebben met onvoorspelbare elementen, kan de statistiek een rol spelen.

N.B.

De meer recente ontwikkelingen in de fysica verlenen statistiek een grotere rol. De kwantum-mechanica gaat er immers van uit dat materie op subatomair niveau zich niet volgens deterministische wetten gedraagt, maar wel met waarschijnlijkheden.

Respons 1.2.6

Een aantal mogelijkheden zijn:

- beschrijven hoe het opleidingsprofiel van de huidige leerkrachtenpopulatie in Vlaanderen eruit ziet;
- beschrijven wat de gemiddelde leeftijd is van de kantoorhouders van banken die deelnemen aan interne bedrijfsopleidingen;

- welk aandeel van de afgestudeerde allochtone studenten uit het hoger onderwijs vindt binnen het jaar een job? Is dit aandeel lager dan bij afgestudeerde autochtone studenten?
- welke opleidingsbehoeften heersen er in de verschillende havenbedrijven in de haven van Antwerpen?
- ...

Respons 1.2.8

Een aantal mogelijkheden zijn:

- verklaren waarom gelijkaardige leerlingen in de ene school veel lager scoren voor wiskunde- en leestoetsen dan leerlingen in een andere school;
- hoe komt het dat de ene bedrijfsopleiding veel meer effect heeft op het effectief gedrag op de werkvloer dan de andere bedrijfsopleiding?;
- waarom is het aandeel van de afgestudeerde studenten uit het hoger onderwijs dat binnen het jaar een job vindt veel lager bij allochtone dan bij autochtone studenten?
- welke structurele kenmerken en welke procesmatige kenmerken van verschillende opleidingen in de verzekeringswereld verklaren de vastgestelde verschillen in effectiviteit van deze opleidingen?
- ...

HOOFDSTUK 2

Het statistisch programma R

DOELSTELLINGEN:

Na dit hoofdstuk

- kan je R en zijn pakketten installeren;
- ken je de gangbare conventies.

2.1. R omgeving



R is een taal en omgeving die de gebruiker in staat stelt statische berekeningen te doen en deze grafisch weer te geven. R is beschikbaar op het internet als ‘General Public License’. Dit komt erop neer dat je het gratis kan downloaden en mag verspreiden. De kracht van R zit, naast de gebruiksvriendelijkheid, in de Open Source. Hierdoor kunnen gebruikers mee schrijven aan programma’s. Dit is echter niet het doel van dit handboek. We beperken ons tot het gebruik van de bestaande programma’s. We spreken van een R omgeving omdat het een volledig programma is. R is dan ook geen afgewerkte onveranderlijk programma maar een veranderlijke omgeving waarbij het basisprogramma wordt aangevuld met pakketten <packages>. Hoe met R en deze pakketten aan de slag te gaan zien we in de volgende hoofdstukken.

2.2. R installeren



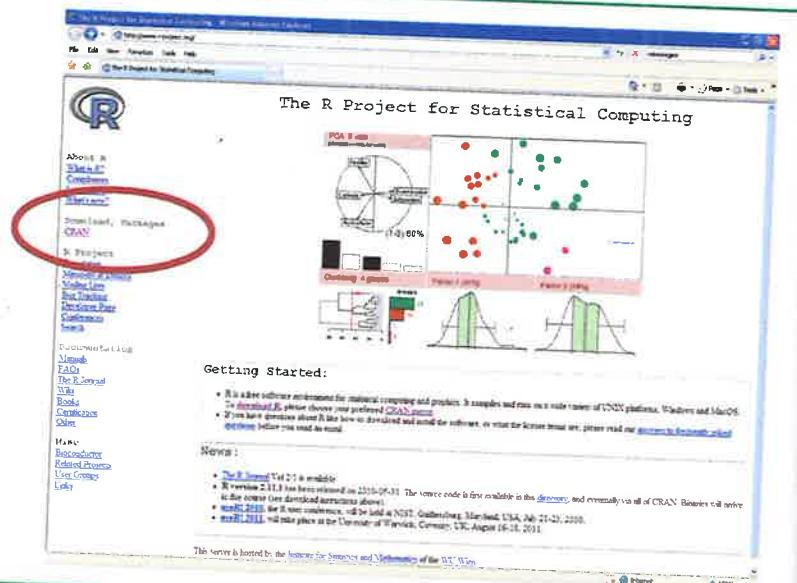
Alvorens met R aan de slag te kunnen moet het programma geïnstalleerd worden op de computer. Hierbij is het ook handig om alvast enkele basis-pakketten te installeren. Dit lichten we toe in enkele stappen. We starten met het beschrijven van de installatie vanuit België. Daarna lichten we kort de installatie in Nederland toe.

INSTALLATIE IN BELGIË:

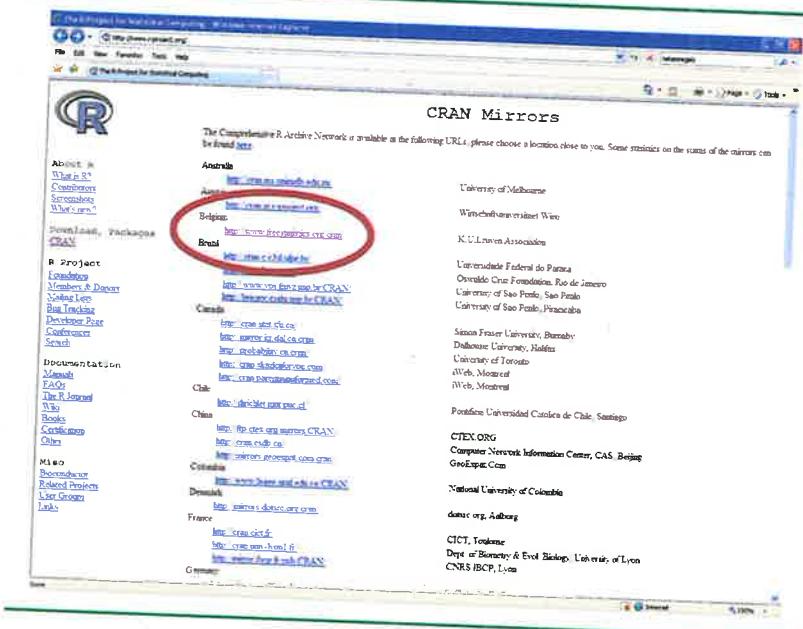
Op de homepage van R (<http://www.r-project.org/>) vind je onder het submenu **Download, Packages** (Fig. 2.1) de link naar de distributieweb-site voor België. (<http://www.freestatistics.org/cran/>) (Fig. 2.2).

R kan geïnstalleerd worden onder Linux, Mac en Windows (Fig. 2.3). In dit handboek beperken we ons tot de installatievoorbeelden van Windows.

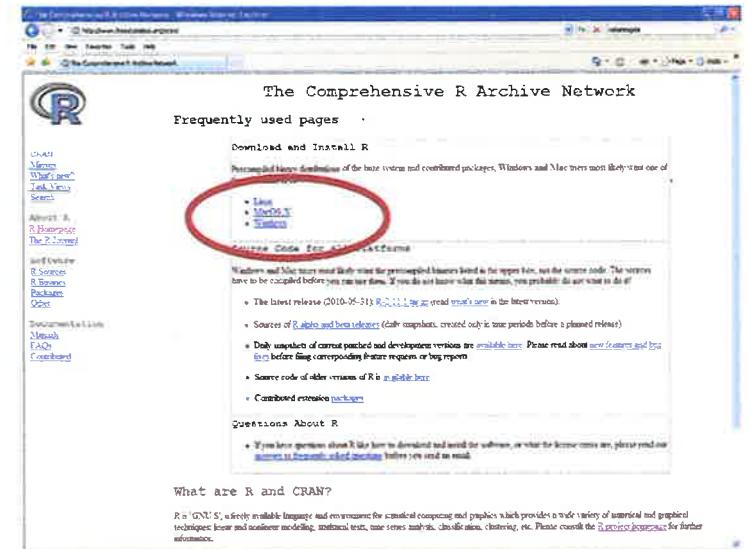
De website van R voorziet eveneens een handleiding en verschillende links om R te installeren.



Figuur 2.1: Screenshot homepage R-project (april 2011)

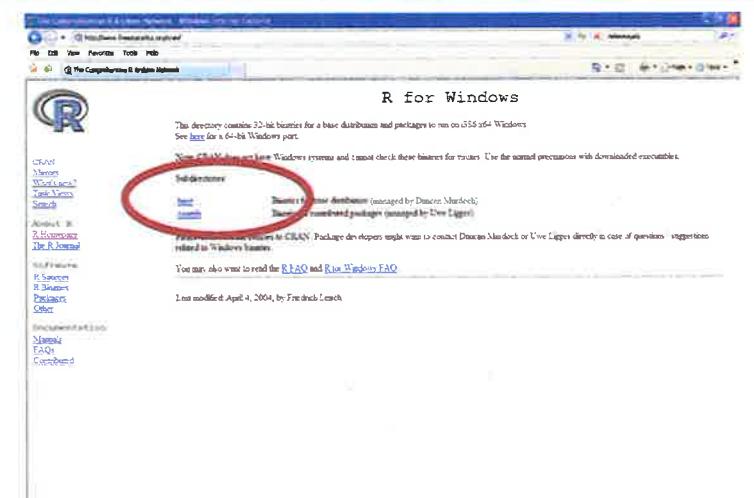


Figuur 2.2: Screenshot homepage R-project (april 2011)

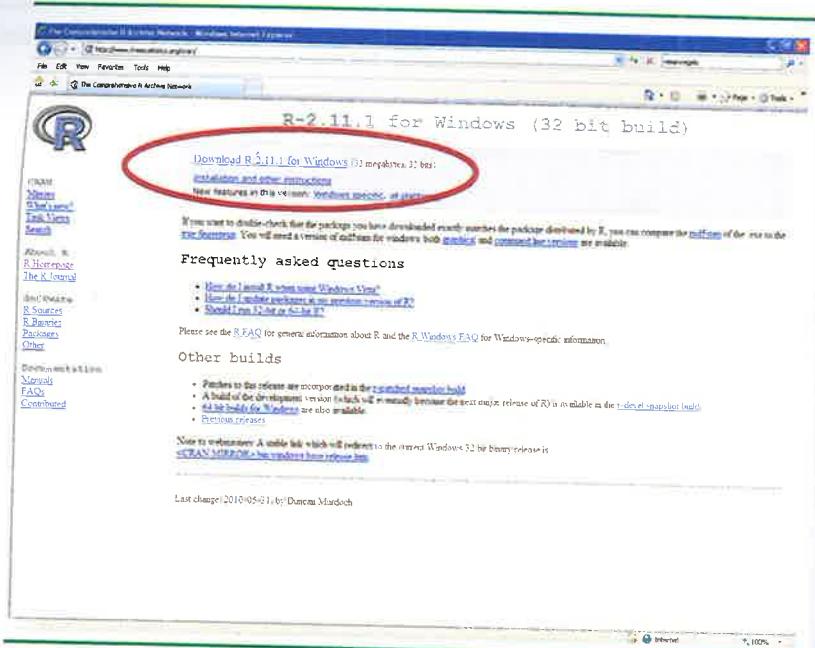


Figuur 2.3: Screenshot homepage R-project (april 2011)

In de meeste gevallen zal je volgende stappen doorlopen.
Klik op "basispakket" Binaries for basispakket distribution (managed by Duncan Murdoch) (Fig. 2.4), vervolgens doorklikken naar *Download R 2.13.0 for Windows* (Fig. 2.5) of recentere uiteraard.



Figuur 2.4: Screenshot homepage R-project (april 2011)



Figuur 2.5: Screenshot homepage R-project (april 2011)

R installeert zich vervolgens zoals de meeste programma's in Windows, waarbij je in enkele dialoogvensters moet aangeven of je wil verder gaan met de installatie.

INSTALLATIE IN NEDERLAND:

Op de homepage van R (<http://www.r-project.org/>) vind je onder het submenu **Download**, **Packages** de link naar de distributiewebsite voor Nederland: <http://cran-mirror.cs.uu.nl/>

Verder is de installatie gelijkaardig aan de Belgische.

Opgelet:

Een beginner installeert best de 32 bit versie. Op sommige computers is het trouwens niet eens mogelijk een 64 bit versie te installeren. Op de website staat echter standaard de 64 bit versie ingesteld. De 32 bit versie vind je terug bij **Other builds**.

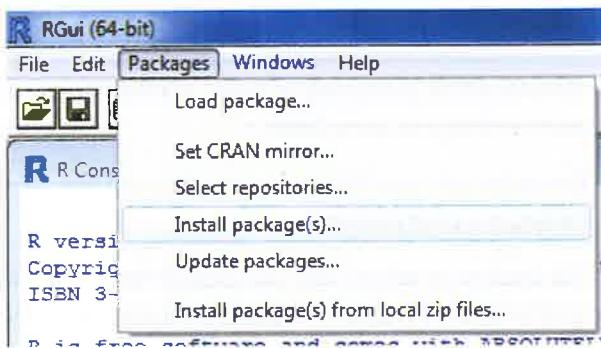
2.3. Installatie van pakketten (packages)

R werkt met pakketten die toelaten bepaalde berekeningen en analyses te doen. Doorgaans zijn deze pakketten ontwikkeld door andere gebruikers om bepaalde functies en analysetechnieken toegankelijker te maken. Behalve de pakketten die standaard voorzien zijn kan je er ook nog zelf installeren. Dit zal de eerste stap zijn.

Na opening van het programma R vind je in het menu **packages** de optie **<install packages...>**.

Vervolgens kies je het land van installatie (België of Nederland).

Er verschijnt een keuzemenu van pakketten (Fig. 2.6).



Figuur 2.6: Screenshot het installeren van pakketten in R

Je kan meteen vaststellen dat hoofdletters en kleine letters een verschil maken. In de bibliotheek van pakketten staan eerst de pakketten met een hoofdletter alfabetisch en vervolgens deze met een kleine letter. Dit heeft niets te maken met de werking van de pakketten, dit is enkel de naam die de maker aan het pakket gaf.

Noot: deze pakketten worden gedownload. Het is dus belangrijk om verbonden te zijn met het internet.

Volgende pakketten zullen in dit boek gebruikt worden en kunnen dus alvast geïnstalleerd worden: "car", "moments".

Je kan ook meerdere pakketten tegelijk inladen door ze te selecteren terwijl je de Ctrl-toets ingedrukt houdt.

Voor gebruikers van MacOs gaat het installeren van pakketten als volgt in zijn werk. Klik in de menubalk op ‘Pakketten en Data’ en kies de optie ‘pakket-installatieprogramma’. De eerste keer dien je het land op te geven van waaruit je de pakketten wilt downloaden. Wanneer er meerdere pakketten ingeladen moeten worden kan dit door de cmd-toets ingedrukt te houden. Vink ook ‘installeer indirect nodige’ aan.

R voorziet voor het installeren van pakketten niet enkel deze ‘aanklikmethode’. Wanneer je liever met functies werkt kan je gebruik maken van de `install.packages` functie. Als voorbeeld om `car` te installeren geef je volgend commando in:

```
> install.packages("car", dependencies=TRUE)
```

2.4. Werken met pakketten ‘*packages*’



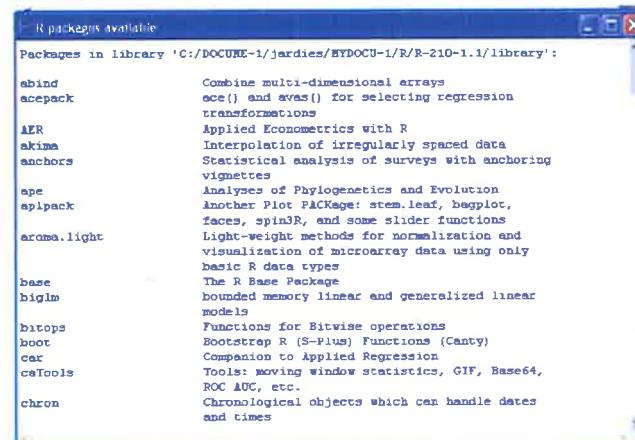
Geïnstalleerde pakketten zijn niet automatisch actief. Om ze actief te maken moet je ze eerst laden.

R voorziet ook voor het laden van pakketten een ‘aanklikmethode’
`Packages > Load package...`

We opteren er echter voor dat je spoedig leert werken met het dialoogvenster van R en zullen uit didactisch standpunt hiervan gebruik maken.

Je kan de pakketten die je reeds ter beschikking hebt, terugvinden in de bibliotheek van pakketten en functies die je in R geïnstalleerd hebt. Deze kan je oproepen door het `library` commando:

```
> library( )
```

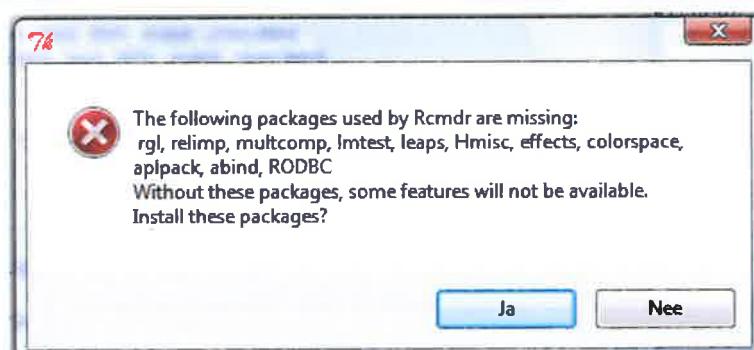


Figuur 2.7: Screenshot van het resultaat van de `library` functie

Van deze functie maak je ook gebruik om een pakket te laden

```
> library("car")
```

Mogelijk ontbreken er, wanneer je een pakket wilt laden, nog pakketten die nodig zijn om het pakket dat jij wilt gebruiken te laten draaien. R zal dit aangeven in een mededeling:



Figuur 2.8: Voorbeeld van een melding van missende pakketten

Om deze bijkomende pakketten te **installeren** en te laden druk je op “Ja”. R haalt dan de pakketten op en installeert deze. Hiervoor is het nodig om met het internet verbonden te zijn.

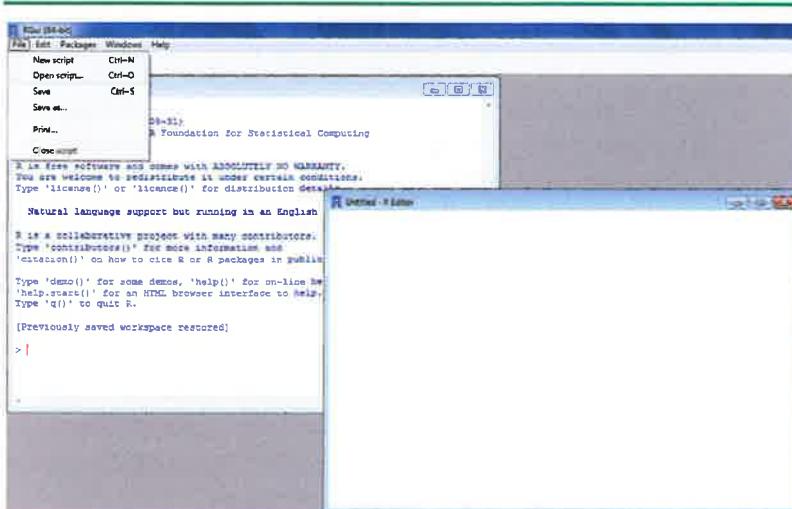
Opmerking: Sommige pakketten bevatten dezelfde functies. Het laatst geladen pakket zal dan deze functies overnemen van het vorige pakket. R geeft dit ook weer wanneer dat gebeurd. (Fig. 2.9)

```
R Console
> library(car)
Loading required package: MASS
Loading required package: nnet
Loading required package: survival
Loading required package: splines

Attaching package: 'car'

The following object(s) are masked _by_ '.GlobalEnv':
      Anova, lht, linearHypothesis
```

Figuur 2.9: Voorbeeld van het overschrijven van functies bij het inladen van een nieuw pakket



Figuur 2.10: Screenshot startscherm R Console en R Editor in Windows

Functies in R, of opdrachten, worden altijd vooraf gegaan door een pijl naar rechts (>) teken. In het 'R console' dialoogvenster (Fig. 2.11) waar de opdrachten zullen uitgevoerd worden staat dit er altijd, gevolgd door de cursor. Het is niet nodig deze pijl opnieuw op te nemen voor de functies in de editor of het scriptvenster.

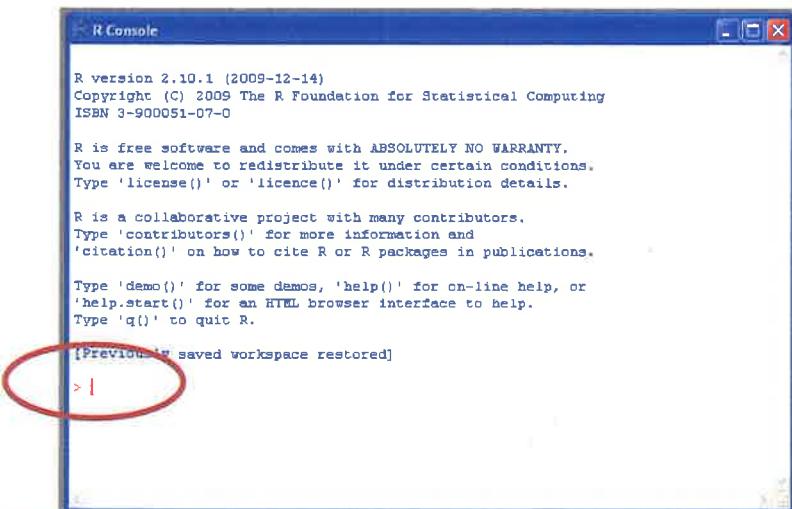
2.5. Conventies in R

2.5.1 Om vlot aan de slag te kunnen gaan met R, is het nodig enkele algemene conventies te leren kennen.



Het is in R meer dan gebruikelijk, zelfs aan te raden, om je functies eerst in een apart scherm te schrijven alvorens ze werkelijk toe te passen. Op deze manier verzamel je verschillende functielijnen bij elkaar. Deze manier van werken laat toe om op een eenvoudige manier eenzelfde functie later opnieuw te gebruiken.

Het nut hiervan wordt tegen het einde van deze cursus meer dan duidelijk. Een afzonderlijk script venster (Fig. 2.10) wordt geopend met de functie 'New Script' in het menu 'File'. In MacOs doe je dit door te klikken op 'Archief' en vervolgens kies je voor 'Nieuw document'.



Figuur 2.11: Screenshot startscherm R Console

Wanneer de functielijn te lang wordt voor één regel wordt ze gesplitst door een + teken toe te voegen op het einde van de eerste regel.

Het is aangewezen om bij complexere statische modellen met vele functielijnen toelichting te schrijven in het script venster. Dit kan op eenvoudige wijze door voorafgaand aan de zin een kardinaalsteken of hekje te plaatsen (#). R herkent dit als een informatieve zin en zal niet trachten hier bewerkingen mee uit te voeren.

Om de lijnen die getypt zijn in de Editor te laten lopen in het scriptvenster volstaat het de cursor in de betreffende lijn te plaatsen en vervolgens Ctrl+R te drukken (de R van ‘run’) of Cmd + Enter in MacOs.

Om meerdere lijnen samen te laten lopen dienen ze samen gemarkeerd te worden en vervolgens eveneens Ctrl +R. Voor MacOS werkt deze functie analoog met door ‘Cmd + Enter’ te drukken.

2.5.2

Specifiek voor deze cursus maken we nog enkele afspraken.



Functies (mogelijke bewerkingen op de data) worden altijd met kleine letters geschreven. Zo zullen we bijvoorbeeld de functie ‘mean’ tegenkomen.

> `mean(Leeftijd)`

Wanneer we zelf een variabele of een datamatrix aanmaken zullen we deze altijd laten beginnen met een hoofdletter gevolgd door kleine letters.

HOOFDSTUK 3

Data en de datamatrix

DOELSTELLINGEN:

Na dit hoofdstuk:

- kan je uitleggen wat we verstaan onder meten in onderzoek;
- kan je uitleggen wat een variabele is;
- ken je de drie sleutelkenmerken van een variabele (totale orde, meeteenheid en een absoluut nulpunt);
- ken je het verschil tussen verschillende soorten variabelen, afhankelijk van het meetniveau;
- kan je een datamatrix opstellen in een rekenblad.



Eén van de bouwstenen voor een statistische analyse is een “variabele”. In dit hoofdstuk staan we stil bij wat we precies onder die term verstaan en de verschillende soorten die we kunnen onderscheiden. Vervolgens lichten we toe hoe we deze individuele variabelen samenbrengen in een data-matrix.

3.1. Wat is data en wat zijn variabelen?

3.1.1

Wetenschap draait rond de confrontatie tussen ideeën (theorieën) en gegevens. Zowel kwantitatief als kwalitatief onderzoek gebruikt dus informatie vervat in gegevens.

3.1.2

We zouden data als volgt kunnen omschrijven:



Data zijn informatie-eenheden die we bekomen hebben via observatie.

Observatie is hier een synoniem voor **meten in de breedste betekenis** van het woord.

3.1.3



Bij welke van de volgende situaties of uitspraken baseren we ons op een meting zoals hierboven bedoeld?

- a) Met de meetlat vaststellen dat een persoon 1m80 groot is;
- b) Vaststellen dat bepaalde personen 1m80 groot zijn of groter en andere personen kleiner dan 1m80;
- c) Jan heeft bruin haar en Vincent is blond;
- d) Uit de verslagen van de vergaderingen van een school leren we dat het aantal fulltime eenheden de afgelopen vijf jaar is gedaald met drie eenheden;
- e) Een gesprek met een buschauffeur leert me dat hij het afgelopen jaar een opleiding heeft gekregen in het omgaan met agressie op z'n bus.

3.1.4



Observeren en daar informatie uit puren en opslaan, dat is de essentie van meten.

Het is belangrijk om hier reeds een onderscheid te maken tussen twee verschillende wijzen van observeren: "open" en "gesloten" observeren. Het onderscheid tussen beide vormen van observatie heeft te maken met het al dan niet op voorhand weten van alle mogelijke "waarden". Stel dat je in een open bevraging aan 100 studenten vraagt hoe ze een bepaalde cursus ervaren hebben, dan spreken we van een "open" observatie. We kennen, zelfs in theorie, niet alle mogelijke antwoorden op die vraag. Vragen we aan dezelfde studenten om de cursus te beoordelen op een cijfer gaande van één tot tien dan spreken we van een "gesloten" observatie: het aantal mogelijke uitkomsten is geweten. Ook de lengte meten in cm's is een gesloten observatie: we kennen in theorie alle mogelijke uitkomsten (ook al zijn die oneindig in aantal). Niemand kan een waarde uitkomen die niet denkbaar is. In dit boek zullen we ons beperken tot het resultaat van gesloten observatie. Op basis van gesloten observatie genereren we wat we noemen een "variabele".

Een **variabele** is een kenmerk van een eenheid uit de populatie **dat op één of andere wijze gemeten kan worden** en **dat varieert** over de eenheden van de populatie heen. Elke variabele heeft een aantal mogelijke gekende waarden. Dit noemen we het domein van de variabele. Dit domein kan op zich oneindig groot zijn.

3.1.5

Wat zijn variabelen en wat niet?



- a) het geslacht van een groep studenten in een welbepaald academiejaar;
- b) de kleur van ogen van dezelfde groep studenten;
- c) het beroep van een groep bankbedienden die een opleiding volgen binnen hun bank;
- d) de onderwijsvorm van een klas leerlingen uit tso-haartooi.

3.2. Het meetniveau van variabelen

3.2.1

Vooraleer we tot de beschrijving en analyse van data overgaan, staan we stil bij hoe meten kan verschillen. Het is van essentieel belang om te weten dat er verschillende soorten metingen bestaan. Allerlei technieken in de statistiek vereisen kennis over de variabelen waarmee je analyseert. Een

verkeerde inschatting van het soort meting kan leiden tot zinloze analyses met bijgevolg nietszeggende resultaten.

3.2.2



Stel, we hebben de twee variabelen lengte en haarkleur gemeten bij vier personen: Vincent, Jan, Peter en Wouter.

- Vincent is 1m70 groot, Peter 1m80, Wouter 1m70 en Jan 1m80
- Vincent en Jan hebben blond haar, Peter bruin haar en Wouter zwart haar

- a) Wat kunnen we allemaal concluderen aangaande de variabele lengte?
- b) Wat kunnen we allemaal concluderen aangaande de variabele haarkleur?
- c) Kan je gelijksoortige conclusies trekken uit beide variabelen?

3.2.3

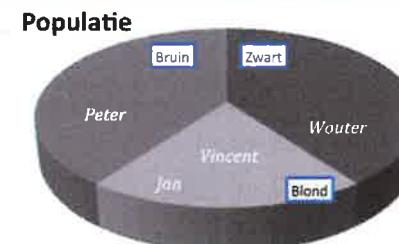


We kunnen meten ook anders definiëren: het indelen van eenheden van een steekproef of populatie volgens één of meerdere kenmerken.

Meerbepaald groeperen we de eenheden die gelijkwaardig zijn voor een bepaald kenmerk. We delen bijvoorbeeld Peter, Jan, Wouter en Vincent als volgt in naargelang hun haarkleur:

- Blond (Vincent – Jan)
- Bruin (Peter)
- Zwart (Wouter)

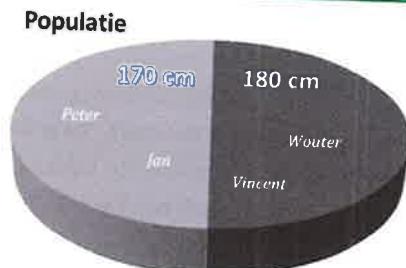
In dit geval hebben we het kenmerk "haarkleur" gemeten en onze eenheden ingedeeld volgens equivalentie klassen. Elke klasse heeft een eigen waarde: de respectievelijke kleur. De onderstaande figuur geeft dit visueel weer:



Figuur 3.1: Verdeling van het kenmerk haarkleur

Een variabele waarbij we enkel kunnen indelen in klassen zullen we **categorische variabelen** noemen. We delen onze eenheden in in categorieën, meer kunnen we niet doen. Een synoniem voor dit type variabelen is **kwalitatieve variabelen**. Zo is de variabele haarkleur een kwalitatieve variabele: blond is niet meer of minder dan bijvoorbeeld bruin als haarkleur. Dit neemt niet weg dat aan de verschillende uitkomsten van kwalitatieve variabelen doorgaans ook cijfers worden toegekend. Zo kunnen we voor de variabele haarkleur een code 1 toekennen aan de categorie 'Blond', code 2 aan de categorie 'Bruin' en code 3 aan de categorie 'Zwart'. Echter, in dit geval hebben de cijfers verder geen enkele betekenis. Ze duiden enkel op een bepaalde categorie. We hadden net zo goed de codes 101, 1001 en 10001 in het leven kunnen roepen. Andere voorbeelden zijn de namen van personen, de woonplaats, het geslacht...

Een meting kan ook **kwantitatieve informatie** opbrengen, zoals bijvoorbeeld bij het meten van de lengte van een persoon. De resulterende categorieën die we onderscheiden zijn getallen die op zich een betekenis hebben. Zo kunnen we een uitkomst 180 bij het meten van een persoon in cm interpreteren omdat er sprake is van een schaal die de getallen op zich betekenis geven. In dat geval spreken we van **kwantitatieve variabelen**. Een synoniem is **numerieke variabele**.



Figuur 3.2: Verdeling van het kenmerk lengte

3.2.4



Stel dat je geïnteresseerd bent in de houding van jongeren ten aanzien van antisociaal gedrag op school.

- Welke kwalitatieve variabele kan je verzinnen om de houding van jongeren ten aanzien van antisociaal gedrag in kaart te brengen?
- Welke kwantitatieve variabele kan je verzinnen om de houding van jongeren ten aanzien van antisociaal gedrag in kaart te brengen?

3.2.5



We hebben een eerste grove indeling in het soort variabelen gemaakt: kwalitatieve versus kwantitatieve variabelen.

De variabelen kunnen echter nog verder onderverdeeld worden aan de hand van 3 cruciale eigenschappen:

- het bestaan van een totale orde;
- het bestaan van een meeteenheid;
- het bestaan van een absoluut nulpunt.

We laten de drie eigenschappen de revue passeren.

(1) HET BESTAAN VAN EEN TOTALE ORDE

Het concept van totale orde kunnen we op een wiskundige wijze definiëren:

- de meetwaarden zijn samenhangend (behoren tot dezelfde hoofdorde). Dit wil zeggen dat ze verwijzen naar dezelfde fenomenen.
- de meetwaarden zijn transitief. Als er een relatie is tussen meetwaarde a en meetwaarde b, en daarnaast is er een relatie tussen meetwaarde b en c, dan is er eveneens een relatie tussen a en c. Bijvoorbeeld: a is meer dan b en b is meer dan c, dan is a eveneens meer dan c.
- de metingen zijn antisymmetrisch: als $a \leq b$ en $b \leq a$, dan is $a = b$.

Iets concreter betekent **totale orde** dat we een rangorde kunnen aanduiden tussen elementen op basis van de meting.

3.2.6



We hernemen onze twee variabelen haarkleur en lengte om dit concreter te maken. Voor welke variabele(n) geldt het principe van het bestaan van een totale orde?

3.2.7



(2) HET BESTAAN VAN EEN MEETEENHEID

Een variabele heeft een meeteenheid als **gelijke verschillen in de waarden** van je variabele, **dezelfde verschillen in intensiteit** van het kenmerk voorstellen.

Als een dergelijke meeteenheid bestaat, dan kan voor elk verschil tussen twee waarden ook worden aangegeven hoeveel meeteenheden het verschil bedraagt.

Een aantal “klassieke” voorbeelden van variabelen die een meeteenheid hebben zijn de variabelen lichaamslengte in cm, temperatuur in °C, gewicht in gram, ... Zo is het verschil tussen 170 en 180 cm even groot als het verschil tussen 180 en 190 cm.

LET OP:

Slechts kwantitatieve variabelen hebben een meeteenheid! Er is niet echt een meeteenheid die bijvoorbeeld het verschil tussen de categorieën blond en bruin kan vatten.

3.2.8



Welke van de volgende variabelen bevatten een meeteenheid?

- a) de houding ten aanzien van pesten op de speelplaats, gemeten aan de hand van een schaal gaande van één tot tien waarbij één staat voor pesten op de speelplaats helemaal niet erg en tien voor dit zeer erg vinden;
- b) het aantal keer dat een jongere zelf iemand gepest heeft op de speelplaats;
- c) zelfbeeld aangaande antisociaal gedrag, gemeten door de jongere zelf een cijfer te laten plakken tussen 1 en 5 voor de mate waarin hij denkt dat anderen hem als een antisociale jongere percipiëren;
- d) het leerjaar waarin een jongere uit het secundair onderwijs zit.

3.2.9



(3) HET BESTAAN VAN EEN ABSOLUUT NULPUNT

Een absoluut of natuurlijk nulpunt is **een waarde (0) die de afwezigheid van het bestudeerde kenmerk weergeeft**.

Dit wordt duidelijk aan de hand van een voorbeeld. Neem de variabele lengte uitgedrukt in centimeters. Als een voorwerp 0cm lengte heeft, dan is het afwezig, dan heeft het geen lengte. De nul is hier een absoluut nulpunt. Een ander typisch voorbeeld zijn aantallen. Het aantal keren dat een kind gepest wordt bijvoorbeeld. Nul wijst bij die variabele op echt nog nooit gepest zijn.

3.2.10



Welke van de volgende variabelen hebben een absoluut nulpunt?

- a) de temperatuur uitgedrukt in graden Celsius;
- b) het gewicht van een boek uitgedrukt in grammen;

- c) het aantal bedrijfsopleidingen dat een werknemer volgde in een jaar;
- d) de IQ-score van een persoon.

3.2.11



Op basis van deze drie kenmerken van variabelen kunnen we de variabelen indelen in vier verschillende soorten. We noemen deze indeling een **indeling van variabelen in meetniveaus**.

Over het algemeen worden de vier types van variabelen of meetschalen als volgt benoemd:

- nominale variabelen
- ordinale variabelen
- intervalvariabelen
- de ratiovariabelen

(1) NOMINAAL MEETNIVEAU

Als we eenheden observeren en van elkaar onderscheiden enkel en alleen gebruik makend van een ‘naam’, dan zitten we op het laagste niveau van meting te observeren. In dat geval wordt er aan geen enkele van de drie eigenschappen voldaan en spreken we van een nominale variabele. Er is geen ordening, er is geen meeteenheid en dus ook geen absoluut nulpunt.

We kunnen dan wel numerieke waarden toekennen aan de eigenschappen die variëren, maar deze waarden doen slechts dienst als index (een code).

Een voorbeeld hiervan is haarkleur:

- 1 → Zwart
- 2 → Bruin
- 3 → Blond

Maar ook: sekse:

- 1 → Man
- 2 → Vrouw

Andere voorbeelden zijn nationaliteit van leerlingen, partijvoorkeur van mensen, onderwijsvorm waarin een leerling zit, sector waarin een arbeider werkt,...

(2) ORDINAAL MEETNIVEAU

Bij een ordinale variabele nemen we aan dat er een ordening aanwezig is (de eerste eigenschap).

Bij meten op nominaal niveau classificeren we de observaties en geven we enkel een naam aan de categorieën. Op ordinaal niveau geven we ranggetallen.

Met andere woorden, ordinale variabelen zijn variabelen die voldoen aan het criterium van totale orde, maar die geen meeteenheid en absoluut nulpunt hebben.

Een voorbeeld hiervan is het toekennen van een score tussen één en tien door een student om aan te geven in welke mate hij pesten op de speelplaats erg vindt. Een negen duidt wel op het erger vinden dan een acht, maar het houdt niet in dat er één punt verschil is in de mate van het erger vinden. Die afstand op zich tussen een negen en een acht heeft geen betekenis.

(3) INTERVAL MEETNIVEAU

Een variabele op intervalniveau vereist het bestaan van een ordening, maar ook van een meeteenheid. Voor elk gevonden verschil (voor elk interval van waarden), kan worden aangegeven in de meeteenheid hoe groot dat verschil (interval) is. Eigen aan intervalvariabelen is dat ze geen absoluut nulpunt bezitten.

Een voorbeeld is de variabele temperatuur uitgedrukt in °C. 1°C is namelijk de hoeveelheid warmte die een kolom kwik van een bepaalde lengte met een afgesproken aantal mm doet toenemen. Tussen 20°C en 30°C liggen tien meeteenheden en tussen 80°C en 90°C liggen eveneens tien meeteenheden. Tussen 60°C en 90°C liggen 30 meeteenheden. Het verschil in temperatuur tussen het interval 60-90 is drie keer zo groot als het verschil in temperatuur tussen het interval 20-30 ($30/10 = 3$).

(4) RATIO MEETNIVEAU

In het geval van een ratiovariabele wordt voldaan aan de drie eigenschappen: er bestaat een totale ordening, er is een meeteenheid aanwezig en daarenboven wordt de afwezigheid van het kenmerk aangeduid door een absoluut nulpunt.

Gemakkelijke voorbeelden zijn aantallen. Bijvoorbeeld het aantal keer dat

een werknemer een opleiding heeft gevolgd. Andere klassieke voorbeelden zijn lengte en gewicht.

We kunnen een duidelijke hiërarchie in de vier verschillende meetniveaus vaststellen. We kunnen ze samenvatten in de volgende tabel:

Tabel 3.1: Samenvatting van de meetniveaus en hun kenmerken

Meetniveau	Nominaal	Ordinaal	Interval	Ratio
Totale orde	-	✓	✓	✓
Meeteenheid	-	-	✓	✓
Absoluut nulpunt	-	-	-	✓

3.2.12



Het vaststellen van type variabele of type meetschaal lijkt evident maar dat is het niet altijd.

Neem het voorbeeld van de manier waarop een leraar cijfers geeft. De leraar geeft een eindbeoordeling op tien.

Wat voor soort variabele is dat volgens jou? Nominaal, ordinaal, interval of ratio?

3.2.13



Nog een voorbeeld. Neem de IQ score van personen.

Wat voor soort variabele is dat volgens jou? Nominaal, ordinaal, interval of ratio?

3.2.14



Ten slotte dienen we mee te geven dat er ook andere indelingen of benamingen bestaan van de soorten variabelen.

Het onderscheid tussen nominaal, ordinaal, interval of ratiovariabelen wordt soms op andere manieren aangeduid. Andere gangbare termen zijn categorische en parametrische of numerieke variabelen.

In de onderstaande tabel beschrijven we hoe de verschillende benamingen zich tot elkaar verhouden:

Tabel 3.2: Meetniveaus en hun andere benamingen

Meetniveau	Stemt overeen met	of met
Nominaal Ordinaal	Kwalitatief	Categorisch

Meetniveau	Stemt overeen met	of met
Interval Ratio	Kwantitatief	Parametrisch/numeriek

Daarnaast wordt in een aantal teksten en onder statistici het onderscheid tussen **discrete en continue variabelen** gehanteerd. Beide termen slaan op variabelen van minstens intervalmeetniveau.

Men spreekt doorgaans van **continue** variabelen als we voor het geobserveerde kenmerk, voor de gemeten variabele, in staat zijn om tussen twee willekeurige meetpunten (waarden) een andere waarde te vinden. Een voorbeeld van een continue variabele is lengte. Tussen 170 en 180 cm vinden we 176 cm. Tussen 170 en 176 vinden we 172. Tussen 171 en 172 vinden we 171,5 cm. En zo kunnen we blijven doorgaan tot een oneindig aantal cijfers na de komma.

Hiertegenover staan zogenaamde **discrete** variabelen: hier vindt men niet altijd een derde waarde tussen twee willekeurige observaties. Voorbeeld van een dergelijke variabele zijn het aantal kinderen in een gezin, het aantal opleidingen dat men volgde per jaar,...

3.2.15 Welke van de volgende gegevens zijn continu en welke discreet?



- a) het aantal aandelen dat elke dag op de beurs wordt verkocht;
- b) temperatuur die elk half uur door het KMI wordt opgemeten;
- c) het jaarlijks inkomen van leraren;
- d) de lengte van 1000 schroeven in een Schroeffabriek;
- e) de toetsscore op een test voor Statistiek.

3.2.16 Het onderscheid tussen soorten variabelen is essentieel in de verdere toepassing van statistische analysetechnieken. Daarom dat we nog een laatste extra voorbeeld meegeven ter verduidelijking van het verschil tussen kwalitatief en kwantitatief:



Wanneer ik aan mijn zoon vraag om een top 5 op te stellen van de volgende bezigheden: Chinees eten, basketbal, tennis, met mama winkelen, de afwas doen, dan geeft hij waarschijnlijk de volgende score: (1) basketbal; (2) tennis; (3) Chinees eten; (4) afwassen; en (5) met mama winkelen.

Intuïtief weet ik als ouder dat hij basketbal en tennis even graag doet, Chinees eten ligt erg dicht in de buurt daarvan en afwassen doet hij wel als je het vraagt zonder tegenpruttelen. Maar winkelen met mama vindt hij altijd verschrikkelijk saai. Dit staat mijlenver verwijderd van de andere scores.

Hij kan bij deze vragen echter geen gelijke scores geven voor de dingen die hij even graag doet (Tennis en Basketbal), net zoals hij aan winkelen met mama geen -100 kan geven.

Hij heeft een rangorde gemaakt, maar er is geen meeteenheid die toelaat de vijf scores onderling te vergelijken. Zijn rangorde kan ook niet vergeleken worden met de rangorde van de andere leden van het gezin op deze vragen. Dochter lief wast helemaal niet graag af en zal dit dan ook zeker op vijf plaatsen. Maar of ze dit even 'niet-graag' doet als de zoon winkelen kunnen we niet afleiden uit deze scoring. We kunnen dus spreken van een ordinale niveau. Meer of minder graag iets doen, is dingen ordenen en dus categorisch of kwalitatief. Maar dit is niet numeriek of kwantitatief want de afstand tussen de eerste en de tweede plaats is niet even groot als tussen de vierde en de vijfde.

3.3.

De datamatrix

3.3.1



Om statistische analyses te kunnen uitvoeren dienen we eerst gegevens te verzamelen. Deze gegevens, of data, zijn niet meer of minder dan stukjes informatie over elke respondent uit je onderzoek. Deze stukjes informatie zijn "vertaald" in meetwaarden voor een bepaalde variabele.

Zo is in de onderstaande tabel informatie over een eerste respondent uitgeengerafeld in verschillende variabelen:

Tabel 3.3: Gegevens van de eerste respondent

Respondent	Naam	Code naam	Seks	Code seks	Leeftijd	Diploma hoger onderwijs	Aantal jaren HO-dipлома behaald
1	Ann	1	Vrouw	1	27	Master	3

Elke kolom bevat een bepaalde variabele. We ontnaaien hierboven de informatie die we optekenden voor deze eerste respondent op 8 variabelen: een nummer die aangeeft over welke respondent het gaat (Respondent); de naam van de respondent (Naam); een code die overeenstemt met de naam (Code naam); het geslacht van de respondent (Sekse); een code die overeenstemt met het geslacht (Code Sekse); de leeftijd (Leeftijd); het diploma dat er behaald werd in het hoger onderwijs (Diploma hoger onderwijs); en het aantal jaren geleden dat de respondent zijn/haar diploma hoger onderwijs behaalde (Aantal jaren HO-diploma behaald).

We kunnen dit doen voor elk van onze respondenten. Zo geven de onderstaande tabellen informatie over drie andere respondenten.

Tabel 3.4 – 3.6: Gegevens van de tweede tot vierde respondent

Respondent	Naam	Code naam	Sekse	Code sekse	Leeftijd	Diploma hoger onderwijs	Aantal jaren HO-diploma behaald
2	Piet	2	Man	2	32	Master	8

Respondent	Naam	Code naam	Sekse	Code sekse	Leeftijd	Diploma hoger onderwijs	Aantal jaren HO-diploma behaald
3	Stef	3	Man	2	23	Bachelor	1

Respondent	Naam	Code naam	Sekse	Code sekse	Leeftijd	Diploma hoger onderwijs	Aantal jaren HO-diploma behaald
4	Ann	1	Vrouw	1	36	Geen	

Van deze stukjes informatie per respondent kunnen we een collage maken. Dit doen we door een **datamatrix** op te stellen.

Tabel 3.7: Datamatrix met vier respondenten

Respondent	Naam	Code naam	Sekse	Code sekse	Leeftijd	Diploma hoger onderwijs	Aantal jaren HO-diploma behaald
1	Ann	1	Vrouw	1	27	Master	3
2	Piet	2	Man	2	32	Master	8
3	Stef	3	Man	2	23	Bachelor	1
4	Ann	1	Vrouw	1	36	Geen	

In een datamatrix plaatsen we de variabelen in de kolommen. Vervolgens creëren we per respondent, of meer generiek per case, een rij. In die rij vullen we de meetwaarden per variabele in voor die case.

3.3.2



Rekenbladen zoals Excel of Calc bieden een eerste oplossing om op systematische wijze een datamatrix te construeren en digitaal bij te houden. In Excel kunnen we de bovenstaande datamatrix opstellen door in de eerste rij de namen in te geven van de variabelen die we meten. Vanaf de 2de rij geven we de meetwaarden per case weer.

Werkmap1							
A	B	C	D	E	F	G	H
1	Respondent	Naam	Code naam	Sekse	Code sekse	Leeftijd	Diploma hoger onderwijs
2							
3	1	Ann	1	Vrouw	1	27	Master
4	2	Piet	2	Man	2	32	Master
5	3	Stef	3	Man	2	23	Bachelor
6	4	Ann	1	Vrouw	1	36	Geen
7							

Figuur 3.3: Voorbeeld van een datamatrix in Excel

Responsen

Respons 3.1.3

In elk van deze situaties wordt er op één of andere manier gemeten. Telkens baseren we ons op een observatie waaruit we informatie halen. Zo halen we ook bijvoorbeeld uit gesprekken informatie. In dit geval maken we niet echt gebruik van een meetinstrument zoals we dat in de klassieke zin van het woord gebruiken (zoals bijvoorbeeld een meter), maar van een andere observatietechniek.

Respons 3.1.5

Geslacht (a) en kleur van ogen (b) zijn variabelen. Beroep (c) en onderwijsvorm (d) zijn geen variabelen: alle bankbedienden zijn bankbedienden van beroep. Er is geen variatie, de eenheden verschillen niet van elkaar wat dit kenmerk betreft. Idem voor de onderwijsvorm van de klas studenten uit tso-haartooi. Dit neemt niet weg dat dezelfde kenmerken in een andere meetsituatie wel een variabele kunnen zijn. Zo kan de onderwijsvorm wel een variabele zijn bij een onderzoek naar de waarden van de leerlingen uit het secundair onderwijs in de laatste graad.

Als de variabele slechts één waarde aanneemt, hebben we niet langer te maken met een variabele maar met een constante.

Respons 3.2.2

a) Met betrekking tot de eerste variabele kunnen we stellen dat Vincent en Wouter even groot zijn, evenals dat Peter en Jan even groot zijn. Ook kunnen we stellen dat de afstand tussen enerzijds Vincent en Wouter en anderzijds Peter en Jan tien cm bedraagt. Ten slotte kunnen we afleiden dat Peter en Jan 1,05 keer zo groot zijn dan Vincent en Wouter.

b) Met betrekking tot de 2^{de} variabele kunnen we enkel stellen dat Jan en Vincent dezelfde haarkleur hebben. Geen van de 3 andere personen heeft dezelfde haarkleur als Peter. Hetzelfde kunnen we ook stellen voor Wouter.

c) Er zit andere informatie vervat in de observaties, we meten iets anders. Maar belangrijker nog, de kwaliteit van de informatie verschilt. Aangaande de lengte kunnen we meer verregaande conclusies trekken dan aangaande de haarkleur. Wat de haarkleur betreft, kunnen we het verschil tussen bijvoorbeeld Vincent en Peter niet vatten in een cijfer.

Respons 3.2.4

a) We zouden een aantal gedragingen kunnen oplijsten en aan de jongeren vragen of ze die al dan niet erg vinden. Je krijgt dan een vraag zoals:

Vraag x. *Geef bij elk van de onderstaande gedragingen aan of je ze al dan niet erg vindt.*

	Ik vind dit erg	Ik vind dit NIET erg
- pesten op speelplaats	<input type="checkbox"/>	<input type="checkbox"/>
- kwaadspreken over iemand op speelplaats	<input type="checkbox"/>	<input type="checkbox"/>
- pesten via internet	<input type="checkbox"/>	<input type="checkbox"/>
- pesten via sms	<input type="checkbox"/>	<input type="checkbox"/>
- vechten op speelplaats	<input type="checkbox"/>	<input type="checkbox"/>
- vechten in klaslokaal	<input type="checkbox"/>	<input type="checkbox"/>
- stelen	<input type="checkbox"/>	<input type="checkbox"/>
- liegen	<input type="checkbox"/>	<input type="checkbox"/>

b) Een numerieke variant zou eruit kunnen bestaan dat je aan de jongeren vraagt om een cijfer te geven tussen 1 en 10 per gedrag, afhankelijk van hoe erg dat ze het vinden. 10 staat daarbij voor 'het zeer erg vinden' en 1 voor 'helemaal niet erg vinden'. In dit geval krijg je aan het einde van de rit cijfers met een zekere betekenis. Hoe hoger het cijfer hoe erger de jongere dat soort gedrag vindt.

Respons 3.2.6

De meting van de variabele lichaamslengte bevat een totale orde:

170 cm < 180 cm. We kunnen de verschillende observaties ordenen, 170cm is kleiner dan 180 cm.

De variabele haarkleur bevat geen totale orde. Het lijkt moeilijk om te besluiten dat zwart ≤ bruin ≤ blond.

We kunnen altijd beslissen om onze variabelen te coderen, bijvoorbeeld:

1 Zwart;

2 Bruin;

3 Blond.

Maar deze codes kunnen niet onderling worden vergeleken. We kunnen geen gebruik maken van hun numerieke eigenschappen: 1 is hier niet "kleiner dan of gelijk aan" 2.

Respons 3.2.8

De variabelen uit b) en d) hebben een duidelijke meeteenheid.

Vijf maal iemand gepest hebben, is één keer meer dan vier maal. Dit verschil blijft even groot als we verder ‘opschuiven’ op de meetschaal. Acht maal iemand gepest hebben blijft één keer meer dan zeven maal iemand gepest hebben.

Ook bij leerjaren gaat dezelfde logica op. Iemand die in het derde leerjaar zit, zit één leerjaar hoger dan iemand die in het tweede leerjaar zit. Dit verschil van één leerjaar blijft even groot indien we iemand die in het zesde leerjaar zit vergelijken met iemand die in het vijfde leerjaar zit.

De variabelen a) en c) hebben geen meeteenheid. We hebben geen reden om aan te nemen dat de ‘afstand’ tussen een score één en twee van een respondent even groot is als het verschil tussen een score twee en een score drie van een respondent. We kunnen de verschillen tussen de scores niet uitdrukken aan de hand van een meeteenheid.

Respons 3.2.10

a) De temperatuur uitgedrukt in graden Celsius: deze variabele heeft geen absoluut nulpunt. 0°C betekent niet de afwezigheid van temperatuur (=energie). Merk op dat temperatuur uitgedrukt in graden Kelvin wel een absoluut nulpunt heeft: 0°K = afwezigheid van Browniaanse beweging = afwezigheid van temperatuur of energie = absoluut nulpunt.

b) Het gewicht van een boek uitgedrukt in grammen: deze variabele heeft wel een absoluut nulpunt. Theoretisch gezien zou een boek nul gram kunnen wegen. Let op, het gaat hier om een theoretisch absoluut nulpunt, maar het kan wel. Een gewicht of temperatuur in $^{\circ}\text{K}$ kan nooit negatief zijn, in tegenstelling van temperatuur uitgedrukt in graden Celsius.

c) Het aantal bedrijfsopleidingen dat een werknemer volgde in een jaar heeft duidelijk een absoluut nulpunt. Iemand kan geen enkele opleiding hebben gevolgd.

d) De IQ-score van een persoon is een variabele zonder absoluut nulpunt. Bepaalt het feit dat een persoon geen enkele vraag uit een IQ-toets juist beantwoordde dat deze persoon geen intelligentie heeft? Het antwoord luidt nee. Mogelijkerwijs hadden we een veel te moeilijke toets voorgelegd.

Respons 3.2.12

Is dit een nominale variabele?

Neen, want er is een duidelijke rangorde, dus de variabele is al zeker een ordinale variabele.

Is dit een intervalvariabele?

Neen, want de afstand tussen 8 en 4 is niet altijd vast. Weet iemand met een 8 twee keer zoveel als iemand met een 4 of weet die 4 keer zoveel als iemand met een 2? Is het verschil tussen 5 en 6 even groot, even belangrijk als het verschil tussen 9 en 10?

Is dit een ratiovariabele?

Op het eerste zicht zou je zeggen van niet, aangezien er geen eenduidige meeteenheid is. Echter deze variabele heeft wel een absoluut nulpunt. 0 punten staat voor niets juist hebben.

Er wordt meestal aangenomen dat het een ratioschaal is, alhoewel we feitelijk kunnen stellen dat het maar een ordinale variabele is.

Respons 3.2.13

Theoretisch verwachten we van een IQ-test dat er een systematische samenhang is tussen de intelligentie van een persoon en de mate waarin hij/zij in staat is om de opgaven op te lossen. Bijgevolg wijst een hogere score op een hogere intelligentie.

Wat als men geen enkele opgave juist oplost? Heeft men dan geen intelligentie? Het antwoord luidt neen. Er is dus geen absoluut nulpunt, en IQ is dus geen ratiovariabele.

Wordt er op intervalniveau gemeten? We kunnen dit nagaan aan de hand van vier fictieve personen:

Persoon A met IQ-score 110

Persoon B met IQ-score 114

Persoon C met IQ-score 40

Persoon D met IQ-score 44

Is het verschil tussen A & B even groot als tussen C & D? Puur cijfermatig zou je zeggen van wel. Echter, het is best mogelijk dat de opgaven die D wist op te lossen in vergelijking tot C gemakkelijker waren, dan diegenen die B wist op te lossen in vergelijking tot A. Bijgevolg kunnen we er niet met 100% zekerheid van uitgaan dat beide intervallen in vastgestelde intelligentie even groot zijn.

Dit brengt ons tot de conclusie dat de variabele in feite ordinal is. Maar indien je de onderzoeks literatuur er op doornemt, zal je zien dat de meerderheid van de onderzoekers doen alsof het een intervalvariabele is.

Respons 3.2.15

- | | | |
|--|---|----------|
| a) het aantal aandelen dat elke dag op de beurs wordt verkocht | = | discreet |
| b) temperatuur die elk half uur door het KMI wordt opgemeten | = | continu |
| c) het jaarlijks inkomen van leraren | = | discreet |
| d) de lengte van 1000 schroeven in een schroeffabriek | = | continu |
| e) de toetsscore op een test voor Statistiek | = | discreet |

HOOFDSTUK 4

Databeheer in R

DOELSTELLINGEN:

Na dit hoofdstuk

- kan je data invoegen in R;
- ken je de twee soorten variabelen die in R worden gehanteerd;
- kan je data bewerken (hercoderen en ermee rekenen) in R;
- kan je data bewaren in R;
- weet je hoe je data uit andere bestandsformaten in R kan inlezen.

NODIGE FILES:

Oefen1.xls

een databestand in Excel dat we zullen hanteren om toe te lichten hoe je een excel-file kan inladen binnen R.

NODIGE PAKETTEN:

car

Dit pakket heet voluit "Companion to Applied Regression analysis" en bevat een aantal zeer handige functies die het werken in R vergemakkelijken.

4.1. Soorten data

4.1.1



In hoofdstuk 3 legden we uit wat variabelen zijn. Bovendien stonden we stil bij de verschillende soorten variabelen en toonden we hoe een “collage” van variabelen samen een datamatrix vormt. Heel dit hoofdstuk staat in teken van hoe we deze data kunnen invoeren en beheren in het statistisch pakket R.

Om aan de slag te kunnen gaan binnen R moeten we weten hoe variabelen in R gedefinieerd worden, welke soorten data ze kunnen bevatten en hoe we data kunnen aanpassen in R.

R is een object-gestuurd programma. Dit houdt in dat alles wat we doen in R kan weggeschreven worden in een object. Een object kan bijvoorbeeld het resultaat bevatten van een berekening. De variabelen en de datamatrix die we zullen hanteren in de analyses worden binnen R weleens beschouwd als types van objecten.

De volgende objecten zijn binnen R gerelateerd aan data en databeheer:

- (data) vectoren, deze kunnen numeriek of categorisch zijn;
- data frames (parallel vectoren).

Een vector is eigenlijk niet meer of niet minder dan een kolom met verschillende waarden in. Deze vectoren zijn bijgevolg de wijze om verschillende variabelen in R te bewaren. Binnen R kunnen we verschillende vectoren combineren met elkaar tot een datamatrix. Deze matrix zullen we in R-termen een “dataframe” noemen.

4.1.2



Vectoren kunnen drie vormen hebben, afhankelijk van het soort van gegevens dat we erin wegschrijven. In wat volgt illustreren we de drie verschillende soorten vectoren met een voorbeeld en maken tegelijk deze vector zelf aan in R. Het aanmaken van vectoren kan via verschillende functies. We gebruiken hier de functie `c()`, welke staat voor ‘concatenate’, waar mee we letterlijk een aaneenschakeling van waarden ingeven.

VECTOR VORM 1: MET CIJFERS ALS GEGEVENS

Een eerste soort vector bevat cijfers als gegevens. Hieronder maken we bij wijze van voorbeeld een nieuwe variabele ‘Leeftijd’ aan voor vijf studenten met de leeftijden 22, 25, 31, 38 & 52.

```
> Leeftijd <- c(22, 25, 31, 38, 52)
```

Indien we in R vervolgens het object Leeftijd (zoals gezegd een vector) oproepen door simpelweg Leeftijd te typen na de prompt, krijgen we zicht op wat dit object bevat. In dit geval onze vijf verschillende leeftijden.

```
> Leeftijd
[1] 22 25 31 38 52
```

VECTOR VORM 2: MET LETTERS ALS GEGEVENS

Naast cijfers kunnen vectoren ook letters, woorden of zinnen bevatten als gegevens. Een voorbeeld is de variabele 'Voornaam'. We maken deze nieuwe variabele aan voor dezelfde vijf studenten. Als we straks deze variabele willen verbinden met de leeftijd van die vijf denkbeeldige studenten, dan is het belangrijk dat we de namen ingeven in dezelfde volgorde als de leeftijden.

```
> Voornaam <- c("Gert", "Liesje", "Anke", "Sven", "Paul")
```

Het maakt niet uit of je hier dubbele of enkele aanhalingstekens gebruikt, maar accenttekens mag niet.

Ter controle:

```
> Voornaam
[1] "Gert" "Liesje" "Anke" "Sven" "Paul"
```

VECTOR VORM 3: LOGISCHE WAARDE ALS GEGEVENS

Als derde en laatste mogelijkheid kunnen we ook nog een logische code gebruiken als waarde voor een variabele. Een waarde kan waar of niet waar zijn (TRUE or FALSE).

Zo kunnen we ingeven of de betreffende studenten werkstudenten zijn of niet.

Stel dat Gert, Liesje en Sven geen werkstudenten zijn en Anke en Paul wel. Dan kunnen we dit als volgt ingeven door een variabele 'Werkstudent' aan te maken waarin een F overeenstemt met geen werkstudent zijn (Werkstudent = FALSE) en een T met wel een werkstudent zijn (Werkstudent = TRUE).

```
> Werkstudent <- c(F, F, T, F, T)
```

Wanneer deze waarden worden opgevraagd wordt dit wel weergegeven als 'TRUE' en 'FALSE'.

```
> Werkstudent
[1] FALSE FALSE TRUE FALSE TRUE
```

Naast dit onderscheid in **vorm**, kunnen vectoren ook van elkaar verschillen in **meetniveau** (zie hoofdstuk 3). Vectoren kunnen gegevens bevatten die op nominaal, ordinaal, interval of rationale niveau gemeten zijn.

4.1.3

Binnen R is er enkel een onderscheid tussen drie types van variabelen: factoren, geordende factoren en numerieke variabelen.

- **Factoren** ('factors') zijn wat we in hoofdstuk 3 onder de koepelterm 'categorische variabelen' hebben geplaatst: nominale en ordinale variabelen.

Een typevoorbeeld hiervan is geslacht. Een ander voorbeeld kan de woonplaats van iemand zijn: Amsterdam, Antwerpen, Breda, Brussel, ...

Binnen de groep van factoren kunnen we een specifieke groep factoren onderscheiden: **geordende factoren**. Dit zijn categorische variabelen waarbij wel sprake is van orde: ordinale variabelen.

Wanneer we dezelfde variabele woonplaats van een orde gaan voorzien, namelijk ordenen op hun populatiegrootte, maken we er een geordende factor van. Bijvoorbeeld: Brussel > Amsterdam > Antwerpen > Breda.

- **Numerieke variabelen** kunnen discreet of continu zijn. Hier hebben we dus te maken met tenminste een zekere ordening.

Numerieke variabele als eigenschap mag niet worden verward met de numerieke vorm. Zo kan in een databestand 1 staan voor "vrouw" en 2 voor "man". De variabele bevat wel numerieke gegevens (m.a.w. er staan cijfers in) maar het is een factor van een nominaal niveau. Er is duidelijk geen sprake van ordening of zijn mannen ook niet 2 keer zoveel als vrouwen.

4.2. Het aanmaken van data

4.2.1

Er bestaan drie soorten functies om variabelen aan te maken in R. In 4.1 gebruikten we reeds de functie `c()` (= 'concatenate') om de verschillende vectoren aan te maken.

Er bestaan nog twee andere functies die soms handig kunnen zijn:

- `seq()` wat staat voor 'sequence' en
- `rep()` wat komt van 'replicate'.

De functie ‘**sequence**’ kunnen we gebruiken om een reeks opeenvolgende cijfers aan te maken. Stel dat we bijvoorbeeld voor elk van onze vijf studenten een uniek nummer willen hebben (een identificatienummer) en dat willen wegschrijven in een variabele ID. Aan de hand van de `seq()` functie maken we deze als volgt aan:

```
> ID <- seq(1,5)
```

Hierbij krijg je als resultaat de reeks cijfers van één tot en met vijf.

```
> ID  
[1] 1 2 3 4 5
```

Meer generiek kunnen we de `seq()` functie als volgt omschrijven: `seq(a, b, c)`. Daarbij staat a voor de laagste waarde en b voor de hoogste waarde. Optioneel kunnen we ook nog in plaats van c een nummer typen wat staat voor de stapgrootte. Standaard, als je geen cijfer in plaats van c typt, worden stappen van telkens de waarde één gebruikt. Het is echter ook mogelijk om in andere stappen te werken. Bijvoorbeeld per half of per twee. Hieronder een voorbeeld waarin we een reeks van getallen definiëren die lopen van vijf tot negen met telkens tussenstapjes van een half punt.

```
> seq(5,9,0.5)  
[1] 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5 9.0
```

De functie ‘**replicate**’ wordt gebruikt om verschillende waarden in te geven aan de hand van herhalingen. Dit kan bijvoorbeeld handig zijn wanneer er codes moeten gegeven worden aan groepen. Stel dat de eerste drie van onze vijf denkbeeldige studenten uit Vlaanderen afkomstig zijn en de overige twee uit Nederland dan kunnen we de volgende functie gebruiken om dit weg te schrijven in een nieuwe variabele ‘Land’ waarbij we code 1 hanteren voor Vlaanderen en 2 voor Nederland. Van de vijf personen zijn de eerste drie Vlamingen en de volgende twee Nederlanders.

Mogelijkheid 1

```
> Landcode <- rep(1:2, c(3,2))  
  
> Landcode  
[1] 1 1 1 2 2
```

Mogelijkheid 2

```
> Landcode <- c(rep(1,3), rep(2,2))  
  
> Landcode  
[1] 1 1 1 2 2
```

In plaats van de verschillende landen een bepaald cijfer mee te geven kan men ze ook een lettercode geven. Lettervormen staan altijd tussen aanhalingstekens.

```
> Landnaam <- rep(c('Vla', 'Ned'), c(3,2))  
  
> Landnaam  
[1] "Vla" "Vla" "Vla" "Ned" "Ned"
```

Het spreekt voor zich dat de functie `replicate` vooral handig is bij grote databestanden. In plaats van voor elke respondent opnieuw eenzelfde gegeven in te voeren kan men dit in grotere gehelen aanpakken. Hierbij is het wel belangrijk dat men exact weet hoeveel respondenten dezelfde waarde moeten krijgen in een bepaalde vector.

4.2.2



Gert heeft een experiment opgezet naar het effect van projectonderwijs binnen statistiek. In twee universiteiten deelt hij de studenten uit eenzelfde studierichting op in twee groepen: experimentele groep en controle groep. Binnen “Universiteit 1” nemen 212 studenten deel aan het onderzoek, binnen “Universiteit 2” zijn dat er 156.

Maak drie vectoren aan die straks de basis vormen voor een dataset.:

- Een vector met daarin een uniek studentnummer voor alle deelnemende studenten per universiteit. Begin bij de tweede universiteit opnieuw vanaf 1 te tellen. Noem deze vector “Id”.
- Een vector met de naam “Univ” met de waarden 1 en 2.
- In elk van de universiteiten zit de helft van de studenten in de controlegroep en de helft in de experimentele groep. Maak een vector aan die aangeeft of een student al dan niet in de experimentele groep zit: een waarde “0” indien niet en een waarde “1” indien wel. Noem deze vector “Exp_groep”.

4.3. Een databestand aanmaken

4.3.1



Naast losse vectoren maakt R ook gebruik van het objecttype ‘**dataframe**’ welke in feite een hele datamatrica bevatt. Om van onze verschillende variabelen een dataframe ‘Studenten’ te maken, volstaat het om de functie `data.frame()` te gebruiken en de variabelen die we wensen in te geven.

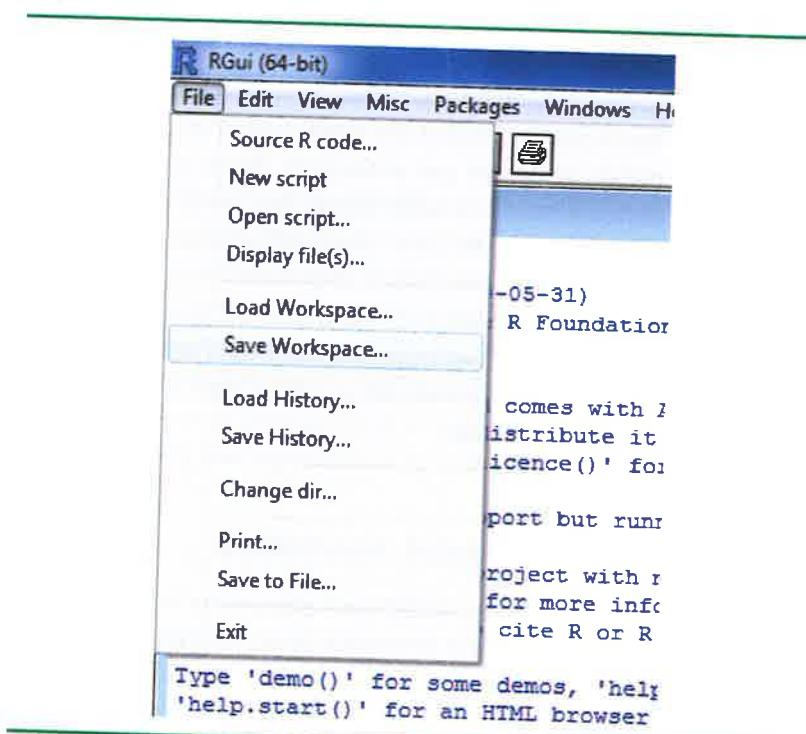
In ons voorbeeld ziet dat er dan als volgt uit:

```
> Studenten <- data.frame(ID, Leeftijd, Voornaam, Werkstudent,
  Landcode)
```

Wanneer we de dataframe willen opvragen in R kunnen we dit zeer eenvoudig doen door het object "Studenten" op te roepen binnen R:

```
> Studenten
   ID   Leeftijd   Voornaam  Werkstudent  Landcode
1  1       22      Gert     FALSE          1
2  2       25    Liesje     FALSE          1
3  3       31     Anke      TRUE          1
4  4       38      Sven     FALSE          2
5  5       52      Paul      TRUE          2
```

Deze dataframe kunnen we vervolgens opslaan met de keuzeknop 'save workspace' in het keuzemenu 'file' in de menubalk. In MacOS klik je daarop 'Workspace' en kies je vervolgens 'Bewaar Workspace bestand'. Een standaard extensie voor datafiles in R is '.RData' of '.rda'.



Figuur 4.1: Het bewaren van een bestand in R

4.3.2



Herneem het voorbeeld van Gert uit 4.2.2. Maak van de drie aangemaakte vectoren een dataframe met de naam "Gert". Bekijk het resultaat door het object "Gert" op te roepen in R.

4.4. Werken in en met een dataframe

4.4.1

Om in ons bestaand dataframe 'Studenten' te werken kan de functie `fix()` gebruikt worden.

```
> fix(Studenten)
```

Dit opent een dialoogvenster met daarin een dataframe met rijen en kolommen (zie figuur 4.2) waarin de verschillende gegevens kunnen aangepast worden.

ID	Leeftijd	Voornaam	Werkstudent	Landcode
1	22	Gert	FALSE	1
2	25	Liesje	FALSE	1
3	31	Anke	TRUE	1
4	38	Sven	FALSE	2
5	52	Paul	TRUE	2

Figuur 4.2: Datamatrix met de functie `fix()`

4.4.2



Wanneer we met een variabele, of meerdere, uit een databestand willen werken moeten we dit aan R kenbaar maken.

Om bijvoorbeeld met de variabele 'Leeftijd' uit de dataframe 'Studenten' te werken, verwijzen we in onze commando's naar het databestand door het voor de variabele te plaatsen gescheiden door een dollar-teken.

```
> Studenten$Leeftijd
```

Wanneer er meerdere bewerkingen in dezelfde dataframe moeten gebeuren is het praktischer om het dataframe als geheel te verbinden met de werkomgeving. Dit gebeurt met het commando `attach()`. Hierna kan de variabele 'Leeftijd' worden opgeroepen zonder dat de dataframe extra vermeld wordt.

```
> attach(Studenten)
> Leeftijd
```

Deze functie blijft geldig tot het dataframe opnieuw losgekoppeld wordt met de functie `detach()`.

Let daarbij op dat je meerdere dataframes tegelijkertijd kan koppelen met de werkomgeving. Dit kan mogelijk tot conflicten leiden bij het aanspreken van variabelen met eenzelfde naam in beide dataframes.

4.4.3



Later zullen we zien dat we de data waarmee we werken vaak moeten gaan "manipuleren". Eén van die vormen van manipuleren is het hercoderen van gegevens. Hercoderen kan in verschillende gevallen nuttig of nodig zijn. Nettig om van cijfers bijvoorbeeld begrijpbare lettercodes te maken. Het hercoderen is in elk geval een nieuwe code geven aan een bestaande code. Soms is dit nodig wanneer er bijvoorbeeld een schaal is die gaat van één tot vijf, maar waarbij vijf de laagste score aangeeft. In dit geval kunnen we de hele schaal omkeren door van een vijf een één te maken, een vier wordt een twee enz...

Om te hercoderen gebruiken we de functie `recode()` uit het pakket `car`. In de meeste gevallen zal dit pakket nog ingeladen moeten worden. Het inladen van pakketten vind je terug in hoofdstuk 2.

```
> library(car)
```

We hernemen de variabele 'Werkstudent' uit 4.1.2. We maakten gebruik van de logische waarden 'True' en 'False'. Deze kunnen we bijvoorbeeld vervangen door 'Ja' en 'Nee'.

Dit kan eenvoudig door het hercoderen van variabelen.

```
> Werkstudent <- recode(Studenten$Werkstudent,
+ "T='Ja' ; F='Nee' ")
```

Hieraan verschijnt in de kolom 'Werkstudent' enkel nog 'Ja' en 'Nee'.

```
> Werkstudent
[1] "Nee"   "Nee"   "Ja"    "Nee"   "Ja"
```

Op bovenstaande manier wijzigen we onze oorspronkelijke variabele. Het is ook mogelijk om een nieuwe variabele aan te maken. Dit kan meteen in één lijn gebeuren.

```
> Studenten$Landnaam <-
  recode(Studenten$Landcode, "1='Vlaanderen' ; 2='Nederland' ")
```

Hiermee is de nieuwe variabele 'Landnaam' meteen aangemaakt in de dataframe 'Studenten' en niet als afzonderlijke vector.

Ter controle bekijken we het overzicht van wat er precies in de dataframe 'Studenten' zit.

```
> Studenten
  ID   Leeftijd  Voornaam  Werkstudent  Landcode  Landnaam
  1   1          22        Gert       FALSE      1         Vlaanderen
  2   2          25        Liesje     FALSE      1         Vlaanderen
  3   3          31        Anke      TRUE       1         Vlaanderen
  4   4          38        Sven      FALSE      2         Nederland
  5   5          52        Paul      TRUE       2         Nederland
```

Merk op dat de variabele 'WerkstudentJN' nog niet aanwezig is. Dit komt omdat (zie hoger) wel een nieuwe variabele aangemaakt hadden, maar deze niet aan het dataframe 'Studenten' hadden toegewezen.

Om dit probleem op te lossen zijn er 2 mogelijkheden.

Optie 1: de variabele opnieuw aanmaken en deze keer meteen toewijzen aan het databestand.

```
> Studenten$WerkstudentJN <- recode(Studenten$Werkstudent,
+ "T='Ja' ; F='Nee' ")
```

Optie 2: De dataframe opnieuw aanmaken en de variabele 'werkstudent' zelf toevoegen.

```
> Studenten <- data.frame(ID, Leeftijd, Voornaam, Werkstudent,
+                           WerkstudentJN, Landcode, Landnaam)
```

Het voordeel van Optie 2 t.o.v. Optie 1 is dat men kan kiezen waar de variabele wordt geplaatst in de dataframe. Bij Optie 1 komt de nieuwe variabele sowieso achteraan.

4.4.4



Het is uiteraard in R ook mogelijk om eenvoudige rekenoefeningen te doen en op die manier nieuwe variabelen aan te maken.

Voorbeeld van deze rekenoefeningen in R:

```
> 2+3
[1] 5
```

Maar het is ook mogelijk om de verschillende variabelen te gebruiken.

We voegen hiervoor de variabele 'Ervaring' toe. Deze geeft aan hoeveel jaren Gert, Liesje, Anke, Sven en Paul al werken.

```
> Ervaring <- c(1,2,5,15,20)
```

Hiermee kunnen we nu bepalen hoeveel % van de totale leeftijd onze respondenten al werken.

```
> Percentwerk <- (Ervaring/Leeftijd)*100
> Percentwerk
[1] 4.545455 8.000000 16.129032 39.473684 38.461538
```

Op deze manier kunnen verschillende eenvoudige bewerkingen worden uitgevoerd waarbij de standaard rekenregels in acht moeten genomen worden.

4.4.5



Eerder in 4.1.3 gaven we aan dat er in R een onderscheid gemaakt wordt tussen "numeric" variabelen en "factor" variabelen. Je kan per variabele in R nagaan wat voor soort variabele het is. Dit doe je door de functie `is.factor()` of de functie `is.numeric()`. Beide functies kan je beschouwen als een vraag die je stelt aan R: `is.factor()` stelt aan R de vraag of de variabele een factor is en `is.numeric()` stelt de vraag of de variabele een numerieke variabele is. We passen dit toe op de variabele Landcode uit Studenten.

```
> is.factor(Studenten$Landcode)
[1] FALSE
> is.numeric(Studenten$Landcode)
[1] TRUE
```

Deze variabele is dus in R als numerieke variabele opgenomen. We kunnen dit aanpassen als we willen door het commando `as.factor()` te hanteren. In het onderstaande commando maken we van de variabele Landcode een categorische variabele.

```
> Studenten$Landcode<-as.factor(Studenten$Landcode)
```

We controleren even of het goed gelukt is:

```
> is.factor(Studenten$Landcode)
[1] TRUE
> Studenten$Landcode
[1] 1 1 1 2 2
Levels: 1 2
```

Uiteraard kunnen we de beweging ook omkeren. We kunnen van een factor ook een numerieke variabele maken met de functie `as.numeric()`. We passen het opnieuw toe op de variabele Landcode:

```
> Studenten$Landcode<-as.numeric(Studenten$Landcode)
> is.numeric(Studenten$Landcode)
[1] TRUE
> Studenten$Landcode
[1] 1 1 1 2 2
```

4.4.6



Mia van de firma Sunseo die zonnepanelen maakt onderzoekt jaarlijks de productiviteit van haar medewerkers. Zo kwam ze aan volgende gegevens over een dagproductie 'Aantalstuks' van haar 15 arbeiders. Deze 15 personen zijn verdeeld over drie verschillende werkposten. Sommigen hebben reeds een vast contract, anderen nog niet. Er zijn mannen en vrouwen op de werkplek. Mannen krijgen de code 0 en vrouwen 1. Om de anonimitet enigszins te bewaken krijgen alle werknemers een nummer van 2010001 tot 2010015 omdat de test in 2010 afgenoem werd.

Het resultaat zal er tenslotte zo moeten uitzien als in figuur 4.3.

Code	Werkpost	Aantalstuks	Vastcontract	Geslacht
2010001	1	23	TRUE	0
2010002	1	24	TRUE	1
2010003	1	15	TRUE	1
2010004	1	16	TRUE	0
2010005	1	15	TRUE	0
2010006	2	13	FALSE	0
2010007	2	25	TRUE	0
2010008	2	15	TRUE	0
2010009	2	18	TRUE	0
2010010	2	19	TRUE	0
2010011	3	15	TRUE	0
2010012	3	13	FALSE	0
2010013	3	19	FALSE	0
2010014	3	18	FALSE	0
2010015	3	24	TRUE	0

Figuur 4.3: Productiviteit werknemers Sunseo

Wanneer Mia met deze cijfers naar de personeelsverantwoordelijke gaat wil ze liever niet dat er nog '0' en '1' staat voor mannen en vrouwen. Je past dit dan ook best nog aan naar 'man' en 'vrouw'.

Het verwachte aantal zonnepanelen dat per dag geproduceerd moet worden per persoon ligt op 22. Geef ook in een nieuwe vector weer hoeveel stuks elke werknemer te veel of te weinig produceert. Dit maakt het voor onze personeelsverantwoordelijke allemaal wat overzichtelijker.

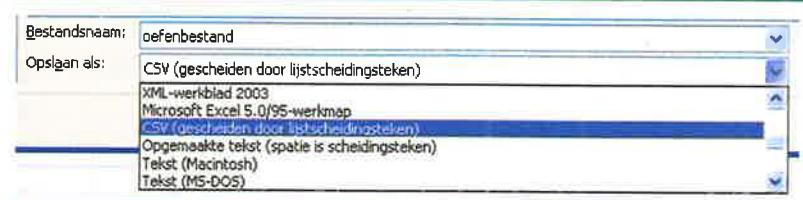
4.5. Bestaande datasets inlezen

4.5.1

 Vaak zullen databestanden reeds beschikbaar zijn in een ander programma dan R. Een van de vaak gebruikte programma's waarin databestanden worden ingevoerd is Excel. Het inladen van data in R vanuit Excel gebeurt als volgt.

Wanneer je het Excel bestand zelf kan bewerken, raden we aan dit bestand op te slaan met de extensie CSV (Comma Separated Variables).

Dit kan via de mogelijkheid "CSV (gescheiden door lijstscheidingstekens)" te kiezen bij "Opslaan als" in Excel.



Figuur 4.4: Het bewaren van een Excel document als CSV document

Vervolgens kunnen we dit bestand in R openen via de functie `file.choose()`, de ruimte tussen de haakjes mag leeggelaten worden, er opent dan een keuzevenster.

```
> file.choose()
```

Om dit bestand meteen een bestandsnaam mee te geven kunnen we opnieuw twee stappen tegelijk zetten.

```
> Oefen1 <- file.choose()
```

Dit bestand moet dan nog worden ingelezen als een tabel. Dit doen we door de functie `read.csv2()`.

```
> Oefen1tabel <- read.csv2(Oefen1)
```

Deze verschillende stappen kunnen echter ook in één stap gezet worden. De commandolijn ziet er dan als volgt uit.

```
> Oefen1tabel <- read.csv2(file.choose())
```

Hiermee opent dan een dialoogvenster waar de betreffende file kan geselecteerd worden.

4.5.2

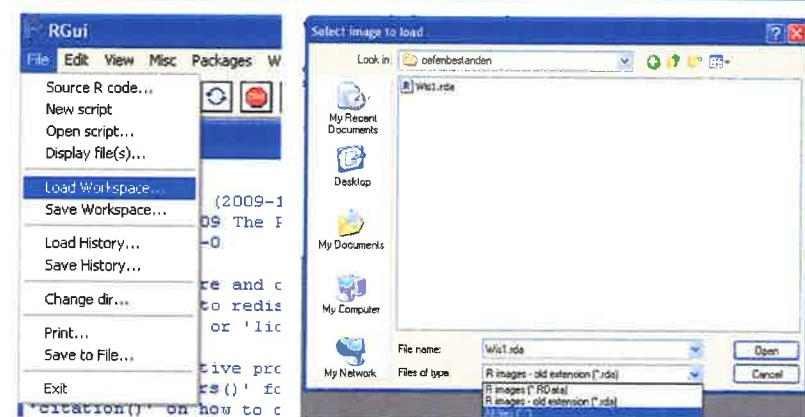


Indien databestanden uit andere statistische softwarepakketten dienen te worden ingelezen, kan gebruik gemaakt worden van het pakket `foreign` in R. Binnen dat pakket kan je verschillende functies terugvinden die je in staat stellen om data in te lezen die afkomstig zijn uit SPSS, SAS, Minitab, Stata, ... We gaan er hier niet verder op in, maar meer informatie daarover is beschikbaar op het net.

4.5.3



Uiteraard is het ook mogelijk om bestaande R bestanden in te lezen. De meest eenvoudige manier om dit te doen is met de functie `load workspace` in het menu `File` (Figuur 4.5a).



Figuur 4.5: (a) Inladen van een bestaand R bestand en (b) selecteren van een bestand

Omdat er verschillende versies van R gebruikt kunnen worden (met de extensie `.RData` of `.rda`) is het raadzaam om te kiezen voor `All files (*.*)` in het keuzemenu filetype (figuur 4.5b).

4.6. Bestaande functies inlezen

Vaak zullen we werken met de file "OLP Functies.R". Deze file bevat een reeks functies die een aantal bewerkingen zullen vereenvoudigen.

Om dergelijke bestaande files met functies in te lezen maken we gebruik van de functie `source()`.

Om dan vervolgens een bepaalde file te selecteren kunnen we gebruik maken van de functie `file.choose()`.

Wanneer we beide functies in elkaar integreren wordt dit:

```
> source(file.choose( ))
```

Responsen

Respon 4.2.2

De eerste vector kan je aanmaken a.d.h.v. het volgende commando.

```
> Id<-c(seq(1,212),seq(1,156))
```

In `seq(1,212)` vragen we om een reeks van getallen te maken gaande van 1 tot 212. Hetzelfde herhalen we van 1 tot 156. Door beide `seq()` commando's te integreren in een `c()` commando gaat R de twee getallenreeksen samenbrengen onder elkaar. Het resultaat kunnen we opvragen door `Id` in te geven in R. Het resultaat hieronder. Merk op dat R in de output niet alles onder elkaar zet, maar naast elkaar. Bij het begin van elke lijn geeft R tussen vierkante haakjes aan bij het hoeveelste element in de vector die nieuwe lijn begint. Zo begint de tweede lijn met het 18^{de} element.

```
> Id
[1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
[18] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
[35] 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
[52] 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68
[69] 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
[86] 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102
[103] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
[120] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
[137] 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153
[154] 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170
[171] 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187
[188] 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204
[205] 205 206 207 208 209 210 211 212 1  2  3  4  5  6  7  8  9
[222] 10  11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
[239] 27  28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
[256] 44  45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
[273] 61  62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
[290] 78  79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94
[307] 95  96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111
[324] 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128
[341] 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145
[358] 146 147 148 149 150 151 152 153 154 155 156
```

Voor de tweede vector kunnen we als volgt te werk gaan:

```
> Univ<-c(rep(1,212),rep(2,156))
```

Bij rep(1, 212) vragen we aan R om het getal 1 212 keer te herhalen. Vervolgens herhalen we het getal 2 156 keer. We brengen dat weer allemaal samen in één vector a.d.h.v. de c() functie. Het resultaat ziet er als volgt uit:

Tot slot dienden we binnen elke universiteit aan te geven of studenten al dan niet in de experimentele groep zaten. Het volgende commando is daar een oplossing voor.

```
> Exp_groep<-c(rep(0,106),rep(1,106),rep(0,78),rep(1,78))
```

We maken wederom gebruik van een combinatie van de `rep()` en de `c()` functie. We herhalen 106 ($=212/2$) keer de waarde 0, vervolgens 106 keer de waarde 1, daarna 78 ($=156/2$) keer opnieuw de waarde 0 en vervolgens 78 keer opnieuw de waarde 1. Hieronder het resultaat:

Respons 4.3.2

Aan de hand van het volgende commando maak je een dataframe aan met de naam Gert. We geven maar een stukje van het resultaat weer (de eerste 10-rijen uit de datamatrix).

```

> Gert<-data.frame(Id,Univ,Exp_groep)
> Gert
    Id     Univ Exp_groep
 1  1       1        0
 2  2       1        0
 3  3       1        0
 4  4       1        0
 5  5       1        0
 6  6       1        0
 7  7       1        0
 8  8       1        0
 9  9       1        0
10 10      1        0

```

Respons 4.4.6

```
# Opnieuw de nodige variabelen (Code, Werkpost, Aantalstuks, Vastcontract,
# Gender en Aantalstukspp) combineren in een dataframe met de naam
# Productiviteit
> Productiviteit = data.frame(Code, Werkpost, Aantalstuks, Vastcontract,
+ Gender,Aantalstukspp)

# Het resultaat bekijken door het object Productiviteit (onze dataframe) op
# te roepen
> Productiviteit

Code      Werkpost    Aantalstuks   Vastcontract   Gender   Aantalstukspp
2010001        1           23       TRUE     man          1
2010002        1           24       TRUE    vrouw         2
2010003        1           15       TRUE    vrouw        -7
2010004        1           16       TRUE     man        -6
2010005        1           15       TRUE     man        -7
2010006        2           13      FALSE     man        -9
2010007        2           25       TRUE     man          3
2010008        2           15       TRUE     man        -7
2010009        2           18       TRUE     man        -4
2010010        2           19       TRUE     man        -3
2010011        3           15       TRUE     man        -7
2010012        3           13      FALSE     man        -9
2010013        3           19      FALSE     man        -3
2010014        3           18      FALSE     man        -4
2010015        3           24       TRUE     man          2
```

Gehanteerde functies

Functie	Doelstelling	Bron
<code>as.factor()</code>	Beschouwt de variabele naar waar verwezen wordt tussen haakjes als een categorische variabele (een factor).	R basispakket
<code>as.numeric()</code>	Beschouwt de variabele naar waar verwezen wordt tussen haakjes als een numerieke (=kwantitatieve) variabele.	R basispakket
<code>c()</code>	Concatenate. Dit commando voegt alle elementen die tussen haakjes zijn geplaatst (gescheiden door een komma) samen in één vector.	R basispakket
<code>data.frame()</code>	Maakt van de vectoren waarnaar verwezen wordt tussen haakjes (gescheiden door een komma) een datamatrix.	R basispakket
<code>file.choose()</code>	Opent een venster waarin je kan "browsen" naar een bestand op de computer waarop je werkt.	R basispakket
<code>is.factor()</code>	Hiermee kan je nagaan of de variabele naar waar je verwijst tussen haakjes in R als een categorische variabele beschouwd wordt.	R basispakket
<code>is.numeric()</code>	Hiermee kan je nagaan of de variabele naar waar je verwijst tussen haakjes in R als een kwantitatieve variabele beschouwd wordt.	R basispakket
<code>rep()</code>	Replicate. Herhaalt een bepaalde waarde meermaals. Het eerste argument tussen haakjes is de waarde die je wil herhalen (kan ook een vector van waarden op zich zijn) en het tweede argument geeft aan hoe vaak die waarde herhaald moet worden. vb.: <code>rep("Guy",10)</code> zal resulteren in 10 keer "Guy".	R basispakket
<code>read.csv2()</code>	Aan de hand van dit commando kan je een Excel file inlezen die opgeslagen is als "csv"-bestand (comma seperated file).	R basispakket
<code>recode()</code>	Dit commando stelt je in staat om een variabele te hercoderen. De specificaties van hoe de variabele gehercodeerd moet worden geef je in het tweede argument tussen dubbele aanhalingstekens. Elke specificatie op zich wordt gescheiden door een punt-komma. vb.: <code>recode(x, "1='ja';2='nee")</code> hercodeert de variabele x als volgt: iedere waarde 1 wordt vervangen door een ja en iedere waarde 2 wordt vervangen door een waarde nee.	car
<code>seq()</code>	Sequence. Dit commando stelt je in staat om een oplopende (of ook aflopende) reeks van getallen te maken. vb.: <code>seq(1:5)</code> geeft een oplopende reeks van 1 tot 5 als resultaat.	R basispakket
<code>source()</code>	Deze functie laat toe om files met functies in te laden.	R basispakket

De frequentieverdeling van een variabele

DOELSTELLINGEN:

Na dit hoofdstuk:

- kan je de termen absolute, relatieve en cumulatieve frequenties uitleggen;
- kan je deze absolute, relatieve en cumulatieve frequenties in R oproepen;
- kan je een frequentietabel lezen, opstellen en oproepen via R;
- kan je na dit hoofdstuk frequentieverdelingen van zowel categorische als kwantitatieve variabelen visueel voorstellen aan de hand van een gepaste figuur in R (histogram, staafdiagram, taartpuntdiagram en puntendiagram).

NODIGE FILES:

Wis1.RData

een file met daarin een aantal variabelen, waaronder de variabelen Score, Iq, Iqcategor en Thuistaal die we doorheen dit hoofdstuk zullen hanteren

OLP Functies.R

een file met daarin aangepaste functies die bij dit OLP horen

 In een groep van 80 leerlingen krijgen we de volgende scores op een wiskundetoets:

Tabel 5.1: Wiskundescores 80 leerlingen

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

We hebben hier te maken met de variabele *wiskundescore*. Als we ons afvragen hoe de meetwaarden van deze variabele verdeeld zijn over deze 80 studenten willen we in feite weten hoe vaak elke uitslag is voorgekomen. Dit soort verdeling zullen we een *frequentieverdeling* noemen. Naast een frequentieverdeling bestaan er ook andere verdelingen: bv. een populatieverdeling, een kansverdeling,... Op dit moment richten we onze aandacht enkel op de frequentieverdeling.

5.1. Absolute en relatieve frequenties

5.1.1

In de beschrijvende statistiek worden verschillende frequenties onderscheiden, zoals onder andere de absolute en de relatieve frequentie.



De **absolute frequentie** is het aantal keer dat een bepaalde meetwaarde of score voorkomt.

Deze absolute frequentie krijgt ook een formele notatieform: n_i .

De n komt van "number" en i geeft aan welke waarde uit de geordende rij van verschillende getallen wordt bedoeld. In Tabel 5.1 zijn er 37 verschillende waarnemingen. Er zijn met andere woorden 37 verschillende meetwaarden waargenomen: 53, 57, 59, 60, 61, 62, 63, 65, 66, 67, 68, 69, 71, 72, enz. De waarde 68 komt overeen met de 11e plaats in die geordende rij, dus $i = 11$. We stellen uit het bovenstaande voorbeeld met de 80 wiskundescores (Tabel 5.1) vast dat de waarde 68 drie maal voorkomt. We kunnen dit noteren als $n_{11} = 3$.

5.1.2



Vul de volgende ontbrekende frequenties in voor de verdeling van de wiskundescores uit Tabel 5.1:

- a) $n_{20} = \dots$
- b) $n_7 = \dots$
- c) $n_{31} = \dots$

- 5.1.3** De absolute frequentie heeft ook een specifieke eigenschap: indien we alle absolute frequenties optellen bekomen we het aantal respondenten.

Dit kunnen we ook aan de hand van een formule weergeven:

$$\sum_{i=1}^p n_i = n$$

De eerste helft van de formule $\sum_{i=1}^p n_i$ leest men als "de som van alle n_i " waarbij i gaat van één tot p ". Daarbij noemt men i de index, één de ondergrens en p de bovengrens van de som. Σ is de Griekse hoofdletter Σ van (Som) en wordt het sommatieteken genoemd.

In Tabel 5.1 zijn er 37 verschillende waarnemingen. Er zijn met andere woorden 37 verschillende meetwaarden waargenomen. In de bovenstaande formule gaat p dus van één tot 37. Passen we de bovenstaande formule toe, dan geeft dit:

$$\sum_{i=1}^{37} n_i = 80$$

- 5.1.4** Het aantal keer dat een bepaalde waarde voorkomt kan men ook uitdrukken in procenten, als een deel van het geheel. In dat geval spreekt men van de **relatieve frequentie** van die bepaalde waarde.

De relatieve frequentie noteren we formeel als: f_i .

- 5.1.5** Als je weet dat n_i staat voor de absolute frequentie van een meetwaarde en n staat voor het totale aantal meetwaarden, hoe zou je dan de relatieve frequentie (f_i) van een meetwaarde in een formule weergeven?

- 5.1.6** Eigen aan relatieve frequenties is:

indien je de som ervan neemt dan is deze gelijk aan één (of 100%).

Ook deze eigenschap is aan de hand van een formule weer te geven:

$$\sum_{i=1}^p f_i = 1$$

$$= \sum_{i=1}^p \frac{n_i}{n} = \left(\frac{n_1}{n} + \frac{n_2}{n} + \frac{n_3}{n} + \dots + \frac{n_p}{n} \right) = \frac{n_1 + n_2 + n_3 + \dots + n_p}{n} = \frac{n}{n} = 1$$

5.1.7



Neem terug de 80 wiskundescores bij de hand uit *Illustratie 5.1*. Vul vervolgens de ontbrekende frequenties in:

- a) $f_{20} = \dots$
- b) $f_8 = \dots$
- c) $f_{16} = \dots$
- d) ... % van de leerlingen behaalt een wiskundescore van 77 punten

5.2. Frequentietabel

5.2.1



Een tabel die de frequentieverdeling van een variabele weergeeft, noemen we een **frequentietabel**.

De tabel bevat per meetwaarde informatie over de absolute en/of de relatieve frequentie. Voor variabelen van ordinaal meetniveau of hoger worden in een frequentietabel de meetwaarden in oplopende volgorde gerangschikt. In tabel 5.2 hernemen we de wiskundescores van de 20 leerlingen uit de eerste twee rijen van tabel 5.1.

Tabel 5.2: Wiskundescores van de eerste 20 leerlingen

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75

In de onderstaande tabel vatten we de frequentieverdeling van deze 20 waarnemingen samen aan de hand van een frequentietabel. Let ook op de opmaak van deze tabel. In publicaties waar de APA-standaard² dient te worden gevolgd, worden tabellen op deze wijze opgemaakt.

In de tabel geef je best eveneens het totaal aantal waarnemingen weer (n). Voor collega-onderzoekers vormt dit essentiële informatie over je resultaten.

-
2. Binnen verschillende disciplines worden verschillende normen gebruikt waaraan een publicatie moet voldoen. Een van die normen is de APA-norm (American Psychological Association) welke in de psychologie gangbaar is. Een onderdeel van die normen is o.a. hoe een tabel (en bijhorende titel) opgemaakt wordt. In R krijg je enkel en alleen de basisgegevens die je dan zelf moet omzetten in een tekstverwerker naar een tabel die aan de opgelegde normen voldoet.

Tabel 5.3: Frequentieverdeling van de wiskundescores van 20 leerlingen

Wiskundescore	Absolute frequentie (n_i)	Relatieve frequentie (f_i)
59,00	1	0,05
60,00	1	0,05
62,00	1	0,05
68,00	2	0,10
71,00	1	0,05
73,00	2	0,10
75,00	2	0,10
76,00	1	0,05
79,00	1	0,05
82,00	1	0,05
84,00	1	0,05
85,00	1	0,05
88,00	2	0,10
90,00	1	0,05
93,00	2	0,10
Totaal (n)	20	1,00

5.2.2



Hieronder vind je de wiskundescores van de leerlingen uit de laatste twee rijen van Tabel 5.1.

Tabel 5.4: Wiskundescores van de laatste 20 leerlingen

65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

- Maak een frequentietabel volgens de APA-normen.
- Hoeveel leerlingen behalen een score hoger dan of gelijk aan 74?
- Wat is het relatief aandeel leerlingen dat lager scoort dan 67?

5.3. Cumulatieve frequenties

5.3.1



Om vragen b) en c) uit puntje 5.2.2 te beantwoorden dien je de absolute of relatieve frequentie op te tellen van alle meetwaarden die hoger of lager zijn dan de gevraagde waarde.

Een handigere wijze om hiermee om te gaan is het toevoegen van de **cumulatieve frequenties** aan de frequentietabel.

De cumulatieve frequentie is het aantal keer dat een bepaalde meetwaarde of een lagere of voorgaande meetwaarde voorkomt in de gegevens.

Het in kaart brengen van cumulatieve frequenties heeft enkel zin bij variabelen van ordinaal meetniveau of hoger.

Cumulatieve frequenties kunnen zowel absoluut (c_i) als relatief (c'_i) zijn. Aan de hand van **absolute cumulatieve frequenties** kunnen we aflezen hoe vaak een bepaalde waarde ‘lager dan of gelijk aan ...’ is. De **relatieve cumulatieve frequenties** geven zicht op hoeveel procent ‘lager dan of gelijk aan ...’ scoort.

In tabel 5.5 geven we weer hoe zowel absolute als relatieve cumulatieve frequenties worden berekend. Hieruit kan je afleiden dat de cumulatieve frequentie van een meetwaarde i in feite gelijk is aan de cumulatieve frequentie van de voorgaande meetwaarde ($i-1$) plus de frequentie van de meetwaarde i zelf (n_i of f_i).

We kunnen dit eveneens samenvatten in een formule:

$$c_i = c_{i-1} + n_i \text{ en } c'_i = c'_{i-1} + f_i$$

Tabel 5.5: Illustratie van wat cumulatieve frequenties zijn

Waarde i	Absolute frequentie (n_i)	Absolute cumulatieve frequentie (c_i)	Relatieve frequentie (f_i)	Relatieve cumulatieve frequentie (c'_i)
1	n_1	n_1	f_1	f_1
2	n_2	$n_1 + n_2$	f_2	$f_1 + f_2$
3	n_3	$n_1 + n_2 + n_3$	f_3	$f_1 + f_2 + f_3$
...
p	n_p	$n_1 + n_2 + n_3 + \dots + n_p$	f_p	$f_1 + f_2 + f_3 + \dots + f_p$

OF

Waarde i	Absolute frequentie (n_i)	Absolute cumulatieve frequentie (c_i)
1	n_1	$c_1 = n_1$
2	n_2	$c_2 = c_1 + n_2$
3	n_3	$c_3 = c_2 + n_3$
...
p	n_p	$c_p = c_{p-1} + n_p$

Een eigenschap van absolute cumulatieve frequenties is dat de cumulatieve frequentie van de hoogste meetwaarde altijd gelijk is aan het aantal waarnemingen (n). De relatieve cumulatieve frequentie van de hoogste meetwaarde is altijd gelijk aan 1.

Dit alles wordt duidelijk uit tabel 5.6. Tabel 5.6 is identiek aan tabel 5.3 met toevoeging van zowel de absolute als relatieve cumulatieve frequentie.

Tabel 5.6: Frequentie, percentage, absolute cumulatieve frequentie en relatieve cumulatieve frequentie van de wiskundesccores van 20 leerlingen

Wiskunde-score	Absolute frequentie (n_i)	Relatieve frequentie (f_i)	Absolute cumulatieve frequentie (c_i)	Relatieve cumulatieve frequentie (c'_i)
59,00	1	0,05	1	0,05
60,00	1	0,05	2	0,10
62,00	1	0,05	3	0,15
68,00	2	0,10	5	0,25
71,00	1	0,05	6	0,30
73,00	2	0,10	8	0,40
75,00	2	0,10	10	0,50
76,00	1	0,05	11	0,55
79,00	1	0,05	12	0,60
82,00	1	0,05	13	0,65
84,00	1	0,05	14	0,70
85,00	1	0,05	15	0,75
88,00	2	0,10	17	0,85
90,00	1	0,05	18	0,70
93,00	2	0,10	20	1,00
Totaal (n)	20	1,00		

De pijlen wijzen op de eigenschap van cumulatieve frequenties.

532

Vul de frequentietabel uit 5.2.2 aan met zowel de absolute en relatieve cumulatieve frequenties (hieronder nogmaals de 20 wiskundescores).



65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

- a) Hoeveel leerlingen behalen een score lager dan of gelijk aan 73?
 - b) Wat is het relatief aandeel leerlingen dat lager scoort dan 66?
 - c) Hoeveel procent van de leerlingen behaalt een score hoger dan 80?

5.3.3



In R kunnen we via verschillende wegen de ingrediënten voor een frequentietabel opvragen. Een eerste eenvoudige stap is het opvragen van de absolute frequenties. Dit kan je via het commando `table()`.

```
> table(wis1$Score)
```

Dit geeft het volgende resultaat:

Ditzelfde resultaat kunnen we wegschrijven in een nieuw object. Dit is handig om de tabel op elk moment opnieuw op te roepen. Een eerste toepassing daarvan is het opvragen van de relatieve frequenties. Om de relatieve frequenties op te vragen kunnen we gebruik maken van het commando `prop.table()`, waarbij `prop` komt van properties.

STAP 1: het resultaat van de functie `table()` wegschrijven in een object met de naam Tabel1

```
> Tabel1<-table(Wis1$Score)
```

STAP 2: gebruik maken van de prop.table() functie

```
> prop.table(Tabell1)
```

Dit geeft het volgende resultaat:

Als we dit liever als percentages weergeven, dan dienen we alles te vermenigvuldigen met 100.

STAP 3: omzetten naar percentages:

```
> prop.table(Tabell1)*100
```

Dit geeft het volgende resultaat:

53	57	59	60	61	62	63	65	66	67	68	69	71
1.25	1.25	1.25	3.75	2.50	5.00	2.50	3.75	1.25	2.50	3.75	1.25	3.75
72	73	74	75	76	77	78	79	80	81	82	83	84
2.50	5.00	3.75	8.75	5.00	2.50	6.25	3.75	1.25	1.25	2.50	1.25	1.25
85	86	87	88	89	90	93	94	95	96	97		
3.75	1.25	1.25	3.75	1.25	1.25	2.50	1.25	2.50	1.25	1.25		

Om echter zowel absolute als relatieve frequenties in één beweging op te vragen en bovendien de cumulatieve frequenties mee op te vragen kunnen we gebruik maken van een functie uit de file "OLP Functies.R": de functie `freqtabel()`.

We doorlopen stap voor stap de procedure om zo'n tabel op te vragen in R.

STAP 1: de file "OLP Functies.R" laden

In een eerste stap gaan we alle functies laden die weggeschreven zijn in de file "OLP Functies.R". De handigste manier daartoe is gebruik maken van het volgende commando:

```
> source(file.choose( ))
```

Dit opent een dialoogvenster uit je besturingssysteem waarin je een file kunt kiezen. Browse naar de locatie waar "OLP Functies.R" staat, selecteer die file en klik op "Open". Vanaf nu kan je gebruik maken van alle functies die in dit bestand staan.

STAP 2: de tabel opvragen

Vervolgens is het vrij simpel om de tabel op te vragen.

```
> freqtabel(Wis1$Score)
```

Het resultaat:

	X	Freq	Percentage	CumulativeN	CumulativePerc
1	53	1	1.25	1	1.25
2	57	1	1.25	2	2.50
3	59	1	1.25	3	3.75
4	60	3	3.75	6	7.50
5	61	2	2.50	8	10.00
6	62	4	5.00	12	15.00
7	63	2	2.50	14	17.50
8	65	3	3.75	17	21.25
9	66	1	1.25	18	22.50
10	67	2	2.50	20	25.00
11	68	3	3.75	23	28.75
12	69	1	1.25	24	30.00
13	71	3	3.75	27	33.75
14	72	2	2.50	29	36.25
15	73	4	5.00	33	41.25
16	74	3	3.75	36	45.00

17	75	7	8.75	43	53.7
18	76	4	5.00	47	58.7
19	77	2	2.50	49	61.2
20	78	5	6.25	54	67.5
21	79	3	3.75	57	71.2
22	80	1	1.25	58	72.5
23	81	1	1.25	59	73.7
24	82	2	2.50	61	76.2
25	83	1	1.25	62	77.5
26	84	1	1.25	63	78.7
27	85	3	3.75	66	82.5
28	86	1	1.25	67	83.7
29	87	1	1.25	68	85.0
30	88	3	3.75	71	88.7
31	89	1	1.25	72	90.0
32	90	1	1.25	73	91.2
33	93	2	2.50	75	93.7
34	94	1	1.25	76	95.0
35	95	2	2.50	78	97.5
36	96	1	1.25	79	98.7
37	97	1	1.25	80	100.0

De kolom met hoofding X bevat de verschillende wiskundescores die voorkomen. Er zijn in totaal 37 verschillende wiskundescores vastgesteld in de variabele (zie eerste kolom). De kolom Freq bevat de absolute frequentie. Onder Percentage staan de relatieve frequenties. De kolom Cumulative bevat de absolute cumulatieve frequenties. De laatste kolom, Cumulative Perc, bevat de relatieve cumulatieve frequenties.

5.3.4



In hetzelfde databestand zit een variabele Iq die de IQ-scores van de leerlingen bevat.

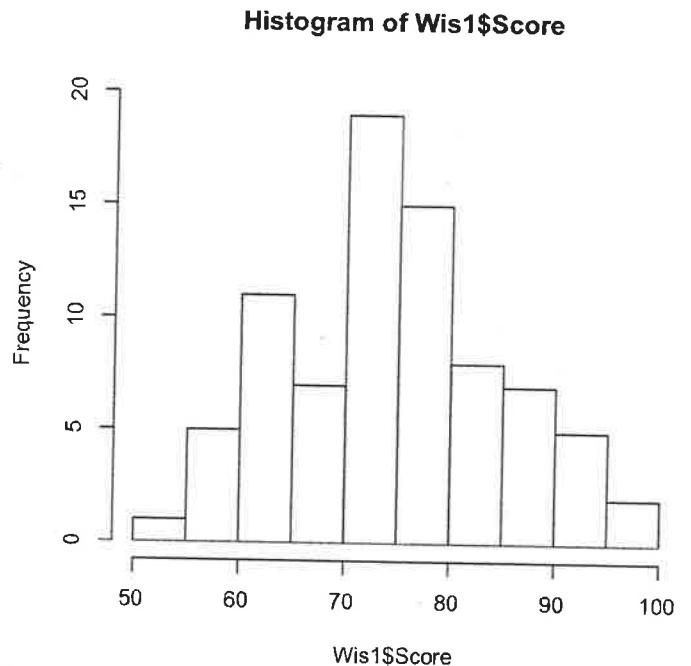
- a) Maak een complete frequentietabel in R voor deze variabele;
- b) Hoeveel verschillende IQ-scores stellen we vast in ons databestand?
- c) Hoeveel leerlingen hebben een IQ-score die gelijk is aan 100?
- d) Hoeveel procent van de leerlingen behaalde een IQ-score van 119,98?
- e) Hoeveel procent van de leerlingen behaalt een score lager dan 90?

5.4. Histogram

5.4.1

 Voor het beschrijven van een frequentieverdeling van een variabele wordt vaak beroep gedaan op een visuele voorstelling ervan. Het visualiseren door middel van een grafiek geeft in vele gevallen een duidelijker indicatie van de verdeling dan louter de cijfers uit een frequentietabel.

Gegevens uit een frequentietabel van een variabele van het interval of ratio meetniveau worden vaak voorgesteld door middel van een **histogram**. Het onderstaand voorbeeld visualiseert de frequentieverdeling van de variabele score uit ons voorbeeld van 80 wiskundescores.



Figuur 5.1: Histogram voor de 80 wiskundescores

Bij het opstellen van een histogram worden de waarnemingen altijd in klassen ingedeeld. Per klasse krijg je een balkje. De breedte van het balkje duidt op de klassenbreedte. In het voorbeeld bedraagt de klassenbreedte vijf punten op de wiskundetoets. De lengte van het balkje geeft de (absolute of relatieve) frequentie aan. Op de Y-as is deze frequentie af te lezen.

Let op: de balkjes raken elkaar. Hiermee maken we duidelijk dat het gaat om een interval- of ratio-variabele: alle meetwaarden op de X-as zijn (theoretisch) mogelijk. Bij sommige variabelen zoals aantallen, is het echter niet mogelijk om minder dan nul te scoren. Maak dit ook duidelijk op het histogram. Geef op de X-as geen minimumwaarden kleiner dan nul weer.

5.4.2



In R kan je vrij snel een histogram van een interval- of ratiovariabele laten maken. Daartoe maak je gebruik van de functie `hist()`. Het voorbeeld in figuur 5.1 is gemaakt via het volgende commando:

```
> hist(Wis1$Score)
```

Uiteraard kan je in R via allerlei bijkomende argumenten het histogram veranderen van ‘looks’. Een eerste aspect dat je mogelijk wil veranderen zijn de labels die worden toegekend aan beide assen. In de onderstaande figuur (5.2) staan verschillende varianten van hetzelfde histogram. Versie (A) bevat door ons aangepaste labels voor beide assen: de x-as hebben we het label *Wiskundescore (Wis1\$Score)* gegeven; de y-as hebben we het label *Absolute frequentie* gegeven. Als hoofdtitel hebben we deze figuur de titel *Histogram van wiskundescores* gegeven.

Dit deden we via het toevoegen van respectievelijk het `xlab=''` argument voor het label van de x-as en het `ylab=''` argument voor het label van de y-as.

De hoofdtitel veranderen we via het argument `main=''`.

Hieronder de aangepaste code, resulterend in versie (A).

```
> hist(Wis1$Score,xlab='Wiskundescore (Wis1$Score)',  
       ylab='Absolute frequentie', main='Histogram van  
       wiskundescores')
```

Een tweede variant (B) bevat in plaats van de absolute frequentie de relatieve frequenties. Daartoe hebben we gebruik gemaakt van het argument `freq=FALSE`. Via dat argument geven we aan dat we geen absolute frequenties willen. De aangepaste code ziet er als volgt uit:

```
> hist(Wis1$Score,xlab='Wiskundescore (Wis1$Score)', ylab=  
       'Relatieve frequentie', main='Histogram van wiskundescores',  
       freq=FALSE)
```

We kunnen ook het aantal balkjes (het aantal klassen) zelf opgeven. Dit is niet altijd aan te raden. R gebruikt namelijk op de achtergrond een formule om het ideale aantal balkjes te berekenen. Maar toch, stel dat we een histogram willen met maar vijf balkjes (zie versie (C)) dan kunnen we gebruik maken van het argument `nclass=`. Hieronder de aangepaste code:

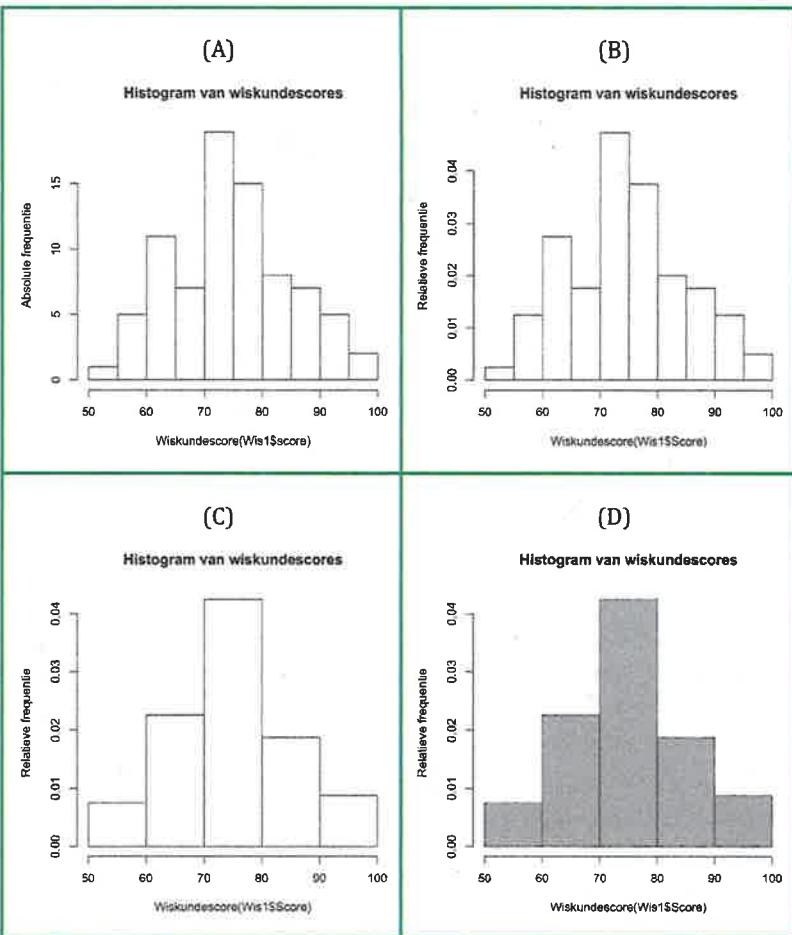
```
> hist(Wis1$Score, xlab='Wiskundescore (Wis1$Score)' ,
       ylab='Relatieve frequentie', main='Histogram van
       wiskundescores', freq=FALSE, nclass=5)
```

Een laatste versie van het histogram dat we presenteren is er eentje waarbij we de balkjes zelf grijs inkleuren. Dit doe je via het argument `col=`. Hier hebben we er concreet voor gekozen om ze grijs in te kleuren. Net zo goed kan je kiezen voor andere kleuren (bv. blue, red, green, yellow,...). De aangevulde code hieronder:

```
> hist(Wis1$Score, xlab='Wiskundescore (Wis1$Score)' ,
       ylab='Relatieve frequentie', main='Histogram van
       wiskundescores', freq=FALSE, nclass=5, col='grey')
```

Tot slot geven we mee dat je eveneens zelf de minimum- en maximumwaarde voor de y-as kan meegeven bij het opmaken van een histogram. Via het argument `ylim=c(y1, y2)` kan je dat doen, waarbij `y1` de minimumwaarde is en `y2` de maximumwaarde. Zo krijg je eigenlijk figuur 5.1, waar we de maximumwaarde op 20 hebben gezet i.p.v. de waarde 15 (vergelijk met figuur 5.2), via het volgende commando:

```
> hist(Wis1$Score, ylim=c(0, 20))
```



Figuur 5.2: Verschillende varianten van het histogram voor de 80 wiskundescores

5.4.3

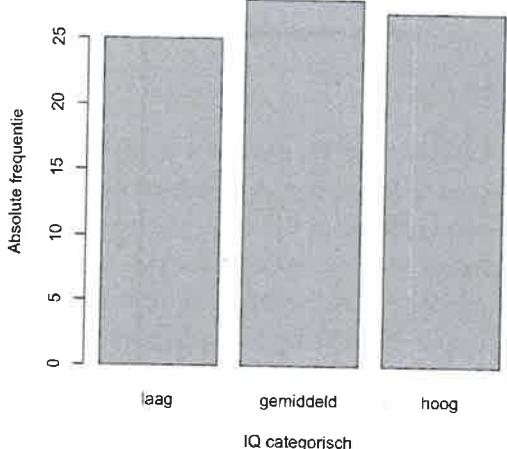
In hetzelfde databestand zitten eveneens 80 IQ-scores.

- a) Maak een standaard histogram voor de variabele iq;
- b) Verander het aantal balkjes, maak een histogram met 4 balkjes;
- c) Verander de y-as: maak er relatieve frequenties van;
- d) Geef zowel de x-as als de y-as een aangepast label + verander de hoofdtitel;
- e) Verander de kleur van de balkjes in blauwe balkjes.

5.5. Grafische voorstellingen van categorische variabelen

- 5.5.1** Tot hiertoe hebben we ons toegespitst op het visualiseren van de verdeling van kwantitatieve variabelen. Om de verdeling over **categorische variabelen** samen te vatten wordt vaak gebruik gemaakt van een **staafdiagram** of een **taartpuntdiagram**. Naast deze twee grafieken zullen we eveneens een **puntendiagram** introduceren, wat vaak een beter alternatief is dan een taartpuntdiagram.

Een **staafdiagram** is sterk vergelijkbaar met een histogram. In een staafdiagram laten we wat ruimte tussen de balkjes in plaats van ze aan elkaar te laten aansluiten. Hieronder een voorbeeld:



Figuur 5.3: Frequentieverdeling van 80 observaties over de drie categorieën van de categorische variant van de variabele IQ heen

De **volgorde van de categorieën** in een staafdiagram is afhankelijk van het soort categorische variabele:

- bij een nominale variabele zijn twee werkwijzen in trek: het **alfabetisch rangschikken** van de categorieën of het **rangschikken in aflopende orde naargelang de frequentie**. De laatste werkwijze geeft meteen meer informatie mee aangezien de lezer van de grafiek in één oogopslag een idee krijgt van welke categorieën het meest voorkomen in de data en welke minder.

- bij een ordinale variabele (zoals in het voorbeeld), worden de categorieën gerangschikt volgens de logische volgorde die in de variabele zelf zit.

5.5.2



Om een staafdiagram aan te maken in R kunnen we wederom meerdere wegen bewandelen. De eerste weg is via het commando `plot()`. Dit is een generiek commando dat voor categorische variabelen ("Factors") een staafdiagram maakt. Toegepast geeft dit:

```
> plot(Wis1$Iqcategor)
```

Om bovendien de x-as en de y-as te voorzien van de juiste labels maken we gebruik van de argumenten `xlab=` en `ylab=`. Passen we dit opnieuw toe, dan krijgen we via het volgende commando de staafdiagram van hierboven:

```
> plot(Wis1$Iqcategor, xlab='IQ categorisch', ylab='Absolute frequentie')
```

Het is echter handiger werken via het commando `barplot()`. Immers, via die weg kunnen we zowel een staafdiagram met absolute als met relatieve frequenties aanmaken als we willen. Het commando `barplot()` verwacht tussen de haakjes geen verwijzing naar een variabele op zich, maar naar het resultaat van een `table()` commando. Hieronder de toepassing om dezelfde figuur van hierboven te maken. Zoals je zal zien is het een beetje omstandiger via die weg. Maar wat verderop zullen we tonen dat het via deze weg ook mogelijk is om een staafdiagram met relatieve frequenties op te roepen.

STAP 1: aanmaken van een object met het resultaat van het `table()` commando

Eerst maken we een object aan met de naam "Freq" (weet dat je hier gerust een eigen naam mag kiezen) waarin we het resultaat van het `table()` commando wegschrijven. In Freq zit nu per categorie opgeslagen wat de bijhorende absolute frequentie is.

```
> Freq<-table(Wis1$Iqcategor)
```

STAP 2: aanmaken van de staafdiagram

Vervolgens dienen we in het commando `barplot()` het object Freq op te roepen. Het commando luidt dan:

```
> barplot(Freq)
```

Weet dat je tevens de x-as en de y-as een eigen label kan meegeven via de argumenten `xlab=` en `ylab=`.

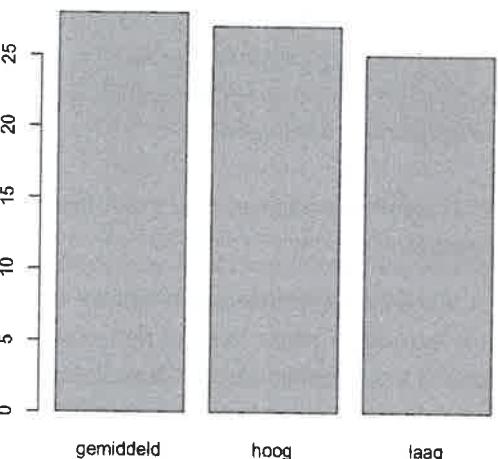
Om een vergelijkbaar staafdiagram aan te maken, maar dan met relatieve frequenties kunnen we zowel `barplot()` als `prop.table()` combineren. Dit ziet er dan zo uit:

```
> barplot(prop.table(Freq))
```

Tot hiertoe hebben we het staafdiagram gereproduceerd zonder aan de volgorde van de balkjes te tornen. Maar zoals we eerder aanhaalden is het voor nominale variabelen vaak handiger om te beginnen met de meest voorkomende categorie en vervolgens de balkjes in aflopende orde te presenteren. Hier zou dit impliceren dat we aan de linkse kant het balkje voor de categorie 'gemiddeld' plaatsen, gevolgd door het balkje van 'hoog' en het balkje van 'laag'. Om dit te bekomen maak je gebruik van de functie `sort()`. Het staafdiagram, met absolute frequenties en aangepaste volgorde verkrijg je via het volgende commando:

```
> barplot(sort(Freq, decreasing=TRUE))
```

Het resultaat:



Figuur 5.4: Frequentieverdeling van 80 observaties overeen de drie categorieën van de categorische variant van de variabele IQ

Let op: deze variabele is in feite een slecht voorbeeld, aangezien deze variabele van ordinaal niveau is. In dat geval behoud je best de natuurlijke

orde in deze variabele als je een staafdiagram aanmaakt. We kozen ervoor om voor deze variabele de volgorde te veranderen omdat op deze wijze duidelijk wordt wat we dienden toe te voegen aan het commando.

5.5.3



Het databestand bevat de variabele Thuistaal. Het betreft een nominale variabele. Maak twee staafdiagrammen aan:

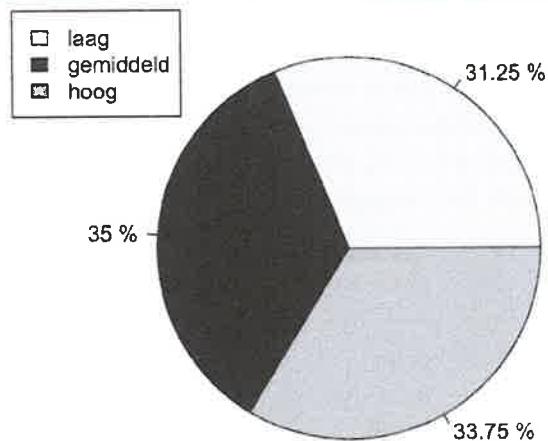
- met absolute frequenties;
- met relatieve frequenties.

Heb telkens ook oog voor de volgorde van de balkjes en voorzie de assen van geschikte labels.

5.5.4



Een vaak gebruikt alternatief voor een staafdiagram is een **taartpuntdiagram** of cirkeldiagram. In een taartpuntdiagram wordt een cirkel opgedeeld in een aantal stukken, één stuk per categorie. De grootte van het stuk staat in relatie tot de frequentie van de categorie. Figuur 5.5 geeft hiervan een voorbeeld:



Figuur 5.5: Frequentieverdeling van de 80 respondenten overeen de drie IQ-categorieën

Dit voorbeeld geeft het resultaat weer van een taartpuntdiagram aangeemaakt in R. Indien je een taartpuntdiagram maakt, dien je zoveel mogelijk informatie mee te geven. In het voorbeeld hebben we per categorie de relatieve frequentie meegegeven.

5.5.

Om een taartpuntdiagram aan te maken in R, inclusief de informatie die we wensen op te nemen (bv. een legende en de relatieve frequenties) moeten we wederom enkele stappen doorlopen. In wat volgt vertrekken we van de standaardfiguur die je krijgt via het `pie()` commando. Vervolgens passen we dit verder aan volgens onze smaak.



STAP 1: aanmaken van een object met het resultaat van een `table()` commando

Net als bij `barplot()` moet je voor dit commando tussen de haakjes verwijzen naar het resultaat van een `table()` commando. We zijn daarbij geïnteresseerd in de variabele `Iqcategor`.

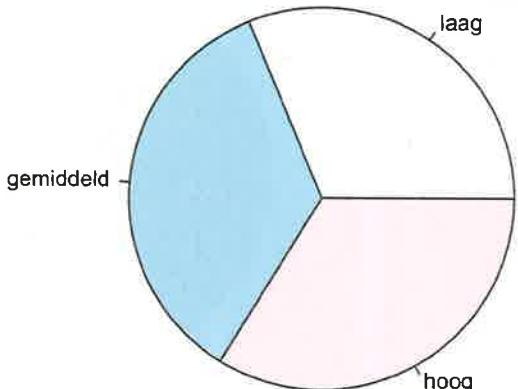
```
> Freq<-table(Wis1$Iqcategor)
```

STAP 2: aanmaken van het taartpuntdiagram

Vervolgens dienen we in het commando `pie()` het object `Freq` op te roepen. Het commando luidt dan:

```
> pie(Freq)
```

Het resultaat is het volgende taartpuntdiagram:



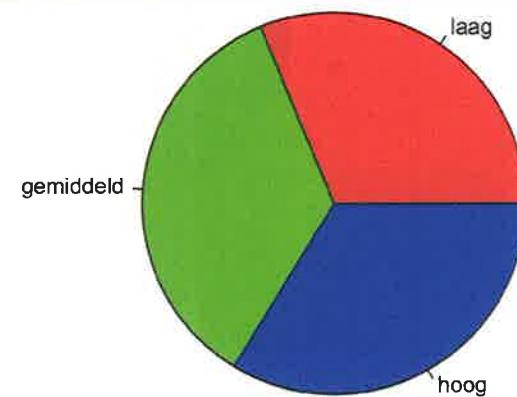
STAP 3: de kleuren aanpassen van de taartpunten

Een van de elementen die we kunnen aanpassen zijn de kleurtjes van de verschillende taartpunten. Dit doe je via het argument `col=` binnen het `pie()` commando. Een voorbeeld maakt dit duidelijk. Via het onderstaande commando vragen we net dezelfde figuur op, maar maken we

gebruik van "regenboogkleuren" in plaats van de pastelkleuren die standaard worden gegeven. Merk op dat we daarbij het aantal categorieën hebben moeten opgeven via `length(Freq)`:

```
> pie(Freq, col=rainbow(length(Freq)))
```

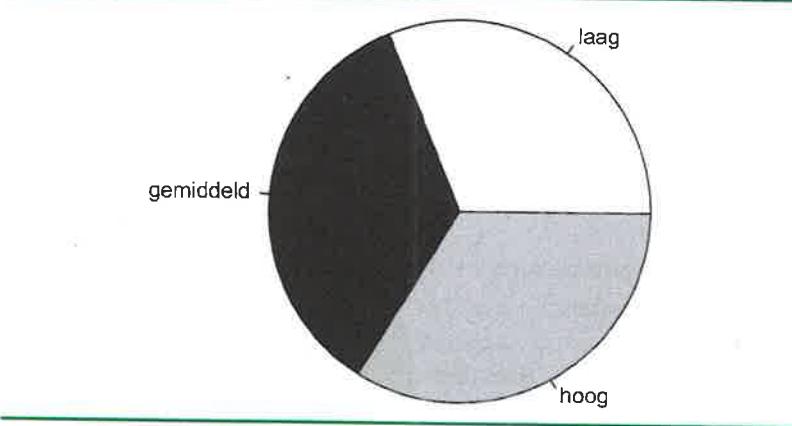
Het resultaat:



Vaak willen we in plaats van gekleurde grafieken gebruik maken van grijswaarden in de grafiek. Hieronder doen we dat in twee stappen. Eerst maken we een nieuw object aan met daarin de drie kleuren die we willen gebruiken voor deze grafiek. We noemen dit object `kleurtjes`. De drie kleuren zijn achtereenvolgens: wit, grijswaarde 20% en grijswaarde 80%. In het `pie()` commando verwijzen we vervolgens naar dit nieuwe object bij het argument `col=`.

```
> kleurtjes<-c('white', 'grey20', 'grey80')
> pie(Freq, col=kleurtjes)
```

Dit is het resultaat:



STAP 4: de labels zelf aanpassen

Nu staat bij elk taartpunt de naam van de categorie als label. Dit kunnen we echter zelf aanpassen via het argument `labels`= in het `pie()` commando. In de voorbeeldfiguur is gebruik gemaakt van de relatieve frequenties in plaats van de namen van de categorieën. Om dit te bekomen zullen we wederom werken met een nieuw object dat we de naam geven 'Eigenlabels' waarin we de relatieve frequenties wegschrijven als labels. Dit kan je op verschillende manieren doen. Wij kiezen ervoor om het `prop.table()` commando te gebruiken.

```
> Eigenlabels <- prop.table(Freq) * 100
```

Het object "Eigenlabels" bevat nu getallen met mogelijk een groot aantal cijfers na de komma. In een figuur is het eleganter om het aantal cijfers na de komma te beperken. Dit kunnen we doen door middel van het commando `round()`. Daarbij geef je als eerste argument welke reeks getallen dienen te worden afggerond en vervolgens geef je aan tot hoeveel cijfers na de komma. Passen we dit toe hier om het aantal cijfers na de komma te beperken tot 2, dan geeft dit het volgend resultaat:

```
> Eigenlabels <- round(Eigenlabels, 2)
```

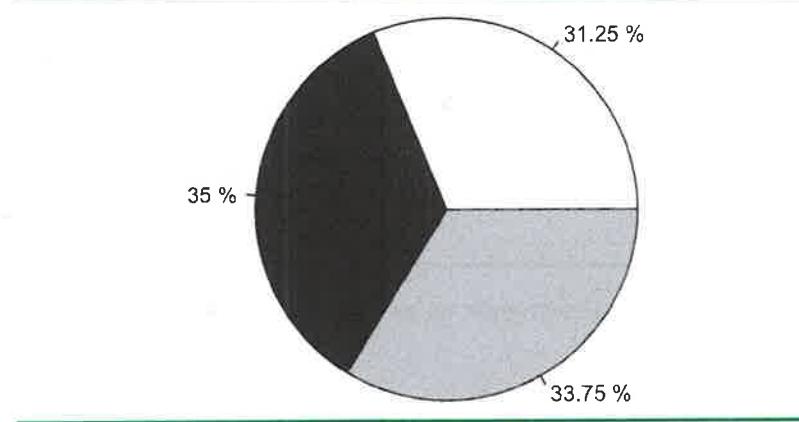
In de figuur zou het vervolgens handig zijn als we duidelijk maken dat de cijfers percentages betreffen. Daartoe kunnen we het % teken toevoegen aan onze eigen labels. Dit doen we via het volgende commando. Dit voegt bij elk getal het %-teken toe, gescheiden door een spatie (`sep=" "`).

```
> Eigenlabels <- paste(Eigenlabels, "%", sep=" ")
```

Laat ons nu het taartpuntdiagram opnieuw aanmaken met daarin de eigen labels:

```
> pie(Freq, col=kleurtjes, labels=Eigenlabels)
```

Het resultaat ziet er zo uit:



STAP 5: een legende toevoegen

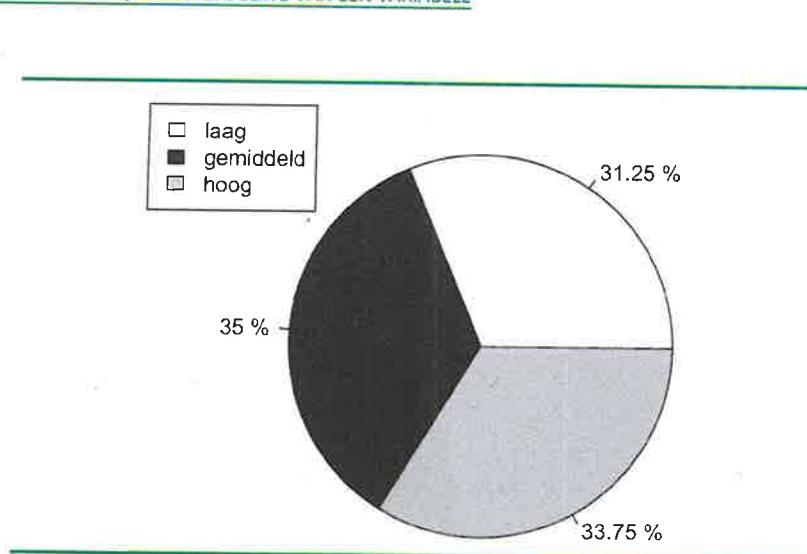
In een laatste stap voegen we een legende toe aan de figuur. Dit doe je door het `pie()` commando te laten volgen door een `legend()` commando. De elementen die we daarbij moeten bepalen zijn achtereenvolgens:

- de horizontale positie (hieronder -1.4);
- de verticale positie (hieronder 1);
- de labels zelf (hieronder `c('laag', 'gemiddeld', 'hoog')`);
- de kleuren in de legende (hieronder `fill=kleurtjes`)

Het hele commando ziet er zo uit:

```
> legend(-1.4, 1, c('laag', 'gemiddeld', 'hoog'), fill=kleurtjes)
```

Na het intypen van dat commando krijgen we de volgende figuur:



Samengebracht geeft dit de volgende commando's om figuur 5.5 na te maken:

```
> Freq <- table(Wis1$Iqcategor)
> kleurtjes <- c('white', 'grey20', 'grey80')
> Eigenlabels <- prop.table(Freq)*100
> Eigenlabels <- round(Eigenlabels,2)
> Eigenlabels <- paste(Eigenlabels,"%", sep=" ")
> pie(Freq, col=kleurtjes, labels=Eigenlabels)
> legend(-1.4,1,c('laag', 'gemiddeld', 'hoog'), fill=kleurtjes)
```

5.5.6



Maak een taartpuntdiagram aan voor de variabele thuistaal in het data-bestand. Zorg ervoor dat dit taartpuntdiagram dezelfde looks heeft als figuur 5.5.

5.5.7

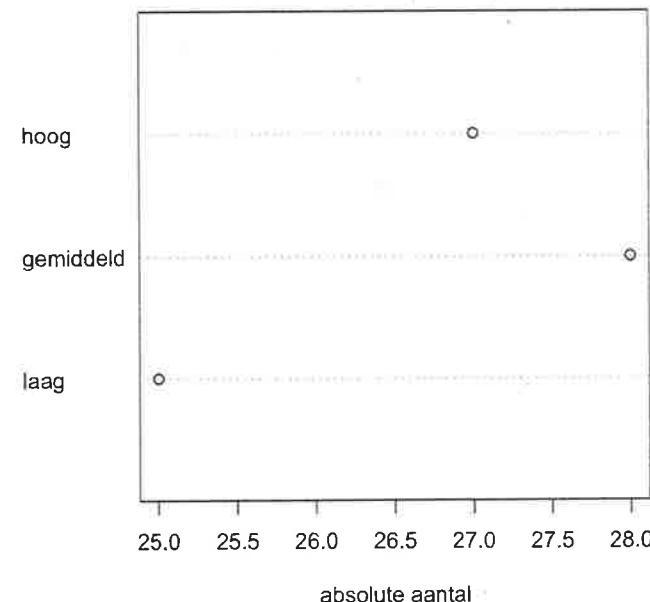


Vanuit de hoek van "perceptuele psychologie" hebben onderzoekers aangegetoond dat een taartpuntdiagram eigenlijk een zeer slechte grafiek is (Cleveland, 1985). De belangrijkste reden die daartoe wordt aangehaald is dat het menselijk brein niet goed in staat is om "hoeken" op het zicht met elkaar te vergelijken.

Cleveland (1985), p. 265:

"Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angel judgements."

Deze dot charts waar hij naar verwijst zullen we **puntendiagrammen** noemen. Figuur 5.6 geeft hiervan een voorbeeld.



Figuur 5.6: Puntendiagram voor de verdeling van de 80 respondenten overeen de drie IQ-categorieën

5.5.8



Een puntendiagram kan je in R maken via het commando `dotchart()`.

STAP 1: aanmaken van een object met het resultaat van een `table()` commando

Net als bij de vorige figuren moet je voor dit commando tussen de haakjes verwijzen naar het resultaat van een `table()` commando. Daarom maken we in een eerste stap een nieuw object aan met de naam `Freq`:

```
> Freq<-table(Wis1$Iqcategor)
```

STAP 2: aanmaken van de puntendiagram

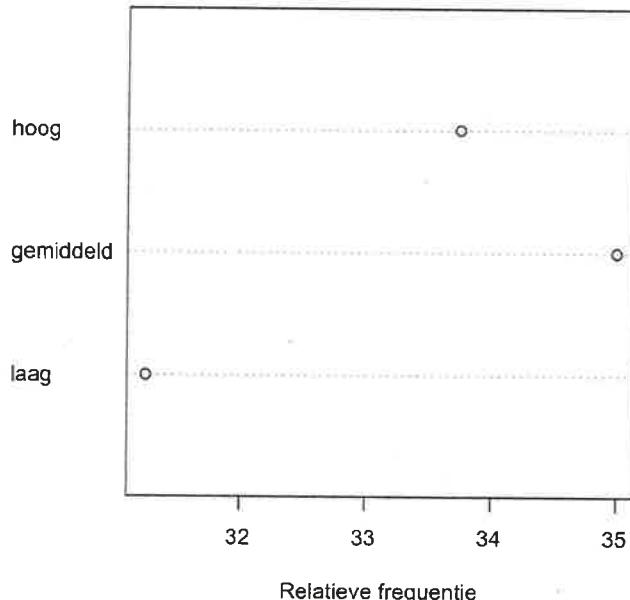
Vervolgens dienen we in het commando `dotchart()` het object `Freq` op te roepen. Het commando luidt dan:

```
> dotchart(Freq)
```

Het resultaat is het volgende puntendiagram uit figuur 5.6, maar dan zonder gepast label voor de x-as. Dit kan je wederom aanpassen door zelf een label te geven via het argument `xlab=" "`. Bovendien kunnen we ervoor opteren om in plaats van absolute frequenties, relatieve frequenties weer te geven. Daartoe maken we gebruik van het `prop.table()` commando dat we kunnen inbedden in het `dotchart()` commando. Hieronder een voorbeeld:

```
> dotchart(prop.table(Freq)*100, xlab="Relatieve frequentie")
```

Het resultaat:

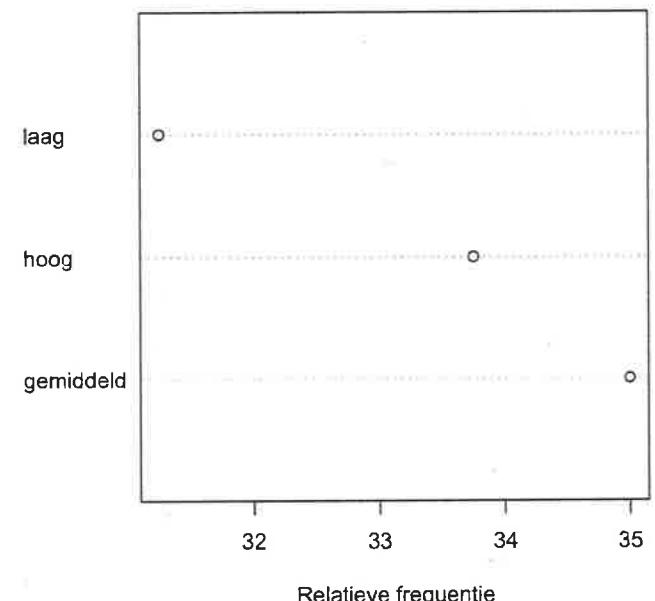


Figuur 5.7: Puntendiagram voor de verdeling van de 80 respondenten overeen de drie IQ-categorieën (in relatieve frequenties)

Ook bij een dotchart kunnen we ervoor opteren om de volgorde van de categorieën op de Y-as te laten afhangen van de frequentie zelf. Hiertoe maken we, naar analogie bij het staafdiagram, gebruik van het `sort()` commando.

Hieronder passen we dit toe (resultaat in figuur 5.8):

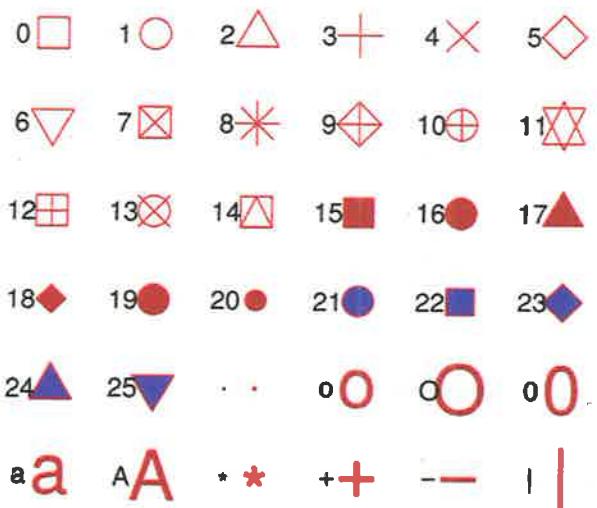
```
> dotchart(sort(prop.table(Freq))*100, decreasing=TRUE, xlab="Relatieve frequentie")
```



Figuur 5.8: Puntendiagram voor de verdeling van de 80 respondenten overeen de drie IQ-categorieën (in relatieve frequenties en gesorteerd)

Een laatste aanpassing die we willen tonen is het zwart maken van de punten in het diagram, om zo deze beter te laten opvallen. Daartoe maken we gebruik van het argument `pch=`. Het argument `pch=` dient om het symbooltype te kiezen. Hieronder staan de verschillende soorten symbooltjes weergegeven in een figuur met hun bijhorend nummer om ze op te roepen in een `pch=` argument. Het open vierkantje komt overeen met nul, voeren we één in dan krijgen we een open bolletje. Twee geeft een open driehoekje en drie resulteert in een plusteken. Zo dien je verder door te tellen om tot de juiste code te komen bij elk symbool. Ook de kleur kan je aanpassen, maar daarover later meer.

**Plot symbols in R;
col = "red3", bg= "slateblue3"**



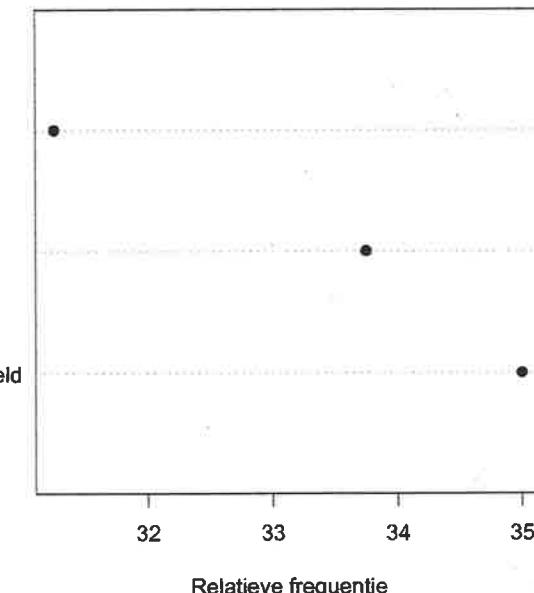
Figuur 5.9: Overzicht van de verschillende waarden voor het pch= argument bij grafieken in R en het bijhorend symbool

Passen we dit toe voor ons puntendiagram dan wordt dit het volledige commando om er eentje aan te maken met relatieve frequentie, gesorteerd in aflopende volgorde en met gevulde bolletjes als symbool (het resultaat in figuur 5.10):

```
> dotchart(sort(prop.table(Freq)*100, decreasing=TRUE),
xlab="Relatieve frequentie", pch=19)
```

5.5.9

Maak een puntendiagram aan voor de variabele Thuistaal in het data-bestand. Geef daarbij de relatieve frequenties weer en sorteer de categorieën in oplopende volgorde op de y-as (bovenaan de meest voorkomende categorie). Als symbool gebruik je een gevuld vierkantje.



Figuur 5.10: Puntendiagram voor de verdeling van de 80 respondenten overeen de drie IQ-categorieën (in relatieve frequenties, gesorteerd en met een ander symbool)

Responsen

Respons 5.1.2

- a) n_{20} = aantal keer waarde 78 = 5
- b) n_7 = aantal keer waarde 63 = 2
- c) n_{31} = aantal keer waarde 89 = 1

Respons 5.1.5

De relatieve frequentie van een meetwaarde wordt berekend door het de absolute frequentie van de meetwaarde (n_i) te delen door het totaal aantal waarnemingen (n). In formulevorm ziet dit er zo uit:

$$f_i = \frac{n_i}{n}$$

Respons 5.1.7

- a) f_{20} = relatieve frequentie van waarde 78 = $5/80 = 6,3\%$
- b) f_8 = relatieve frequentie van waarde 65 = $3/80 = 3,8\%$
- c) f_{16} = relatieve frequentie van waarde 74 = $3/80 = 3,8\%$
- d) 2,5 % van de leerlingen behaalt een wiskundescore van 77 punten



Respons 5.2.2

- a) Een frequentietabel volgens APA-normen zou er zo moeten uitzien:

Tabel 5.2.2: Frequentieverdeling van de wiskundescores van 20 leerlingen

	Frequentie (n_i)	Percentage (f_i)
53,00	1	5,0
57,00	1	5,0
62,00	1	5,0
63,00	1	5,0
65,00	1	5,0
67,00	1	5,0
72,00	1	5,0
73,00	2	10,0
74,00	1	5,0
75,00	1	5,0
76,00	2	10,0
77,00	1	5,0
78,00	1	5,0
80,00	1	5,0
81,00	1	5,0
85,00	1	5,0
86,00	1	5,0
88,00	1	5,0
Totaal (n)	20	100,0

Wat niet mag ontbreken zijn:

- een tabelnummer;
- een titel;
- titels voor de kolommen;
- een rij die aangeeft hoeveel waarnemingen er in totaal waren;

De opmaak dient conform te zijn aan die van de bovenstaande tabel.

- b) Er zijn in totaal 11 leerlingen die hoger dan of gelijk aan 74 scoren. In de tabel kunnen we dit gemakkelijk bekomen door de som te nemen van alle absolute frequenties voor de waarnemingen hoger dan of gelijk aan 74.
- c) 25% van de leerlingen behaalt een wiskundescore lager dan 67

Respons 5.3.2

De frequentietabel aangevuld met cumulatieve frequenties zou er zo moeten uitzien:

Tabel 5.3.2: Frequentieverdeling van de wiskundescores van 20 leerlingen

	Frequentie (n_i)	Percentage (f_i)	Absolute cumulatieve frequentie	Relatieve cumulatieve frequentie
53,00	1	5,0	1	5,0
57,00	1	5,0	2	10,0
62,00	1	5,0	3	15,0
63,00	1	5,0	4	20,0
65,00	1	5,0	5	25,0
67,00	1	5,0	6	30,0
72,00	1	5,0	7	35,0
73,00	2	10,0	9	45,0
74,00	1	5,0	10	50,0
75,00	1	5,0	11	55,0
76,00	2	10,0	13	65,0
77,00	1	5,0	14	70,0
78,00	1	5,0	15	75,0
80,00	1	5,0	16	80,0
81,00	1	5,0	17	85,0
85,00	1	5,0	18	90,0
86,00	1	5,0	19	95,0
88,00	1	5,0	20	100,0
Totaal (n)	20	100,0		

- a) De absolute cumulatieve frequentie van de meetwaarde 73 bedraagt 9. Met andere woorden, negen leerlingen halen een wiskundescore van 73 of lager.
- b) De relatieve cumulatieve frequentie van de meetwaarde 66 is niet af te lezen in de tabel. Wel kunnen we de relatieve cumulatieve frequentie aflezen van de meetwaarde 65. Deze bedraagt 25. Bijgevolg kunnen we afleiden dat 25% van de leerlingen een wiskundescore behaalt die lager is dan 66.
- c) Dit bekom je door de volgende formule tot te passen: $100 - \text{relatieve cumulatieve frequentie van de meetwaarde } 80$. In dit geval: $100-80=20$. Met andere woorden, 20% van de leerlingen behaalt een wiskundescore die hoger is dan 80.

Respons 5.3.4

a) Het volgende commando had je nodig:

> freqtabel (Wis1\$Iq)

De output uit R ziet er zo uit:

	X	Freq	Percentage	CummulativeN	CummulativePerc
1	68.51	1	1.25	1	1.25
2	69.48	1	1.25	2	2.50
3	71.42	1	1.25	3	3.75
4	72.39	1	1.25	4	5.00
5	76.28	1	1.25	5	6.25
6	78.22	1	1.25	6	7.50
7	80.16	1	1.25	7	8.75
8	81.13	1	1.25	8	10.00
9	83.08	1	1.25	9	11.25
10	84.05	2	2.50	11	13.75
11	85.99	1	1.25	12	15.00
12	86.96	1	1.25	13	16.25
13	88.9	3	3.75	16	20.00
14	89.87	5	6.25	21	26.25
15	90.85	1	1.25	22	27.50
16	90.96	1	1.25	23	28.75
17	91.82	2	2.50	25	31.25
18	92.79	5	6.25	30	37.50
19	93.76	1	1.25	31	38.75
20	94.73	3	3.75	34	42.50
21	95.7	4	5.00	38	47.50
22	96.67	3	3.75	41	51.25
23	97.64	2	2.50	43	53.75
24	98.62	2	2.50	45	56.25
25	99.59	1	1.25	46	57.50
26	100.56	2	2.50	48	60.00
27	101.53	1	1.25	49	61.25
28	102.5	1	1.25	50	62.50
29	103.47	2	2.50	52	65.00
30	104.44	1	1.25	53	66.25
31	105.41	1	1.25	54	67.50
32	106.39	1	1.25	55	68.75
33	107.36	1	1.25	56	70.00
34	108.33	1	1.25	57	71.25
35	109.3	2	2.50	59	73.75
36	113.19	1	1.25	60	75.00
37	115.13	2	2.50	62	77.50
38	116.1	2	2.50	64	80.00
39	117.93	1	1.25	65	81.25
40	118.04	2	2.50	67	83.75

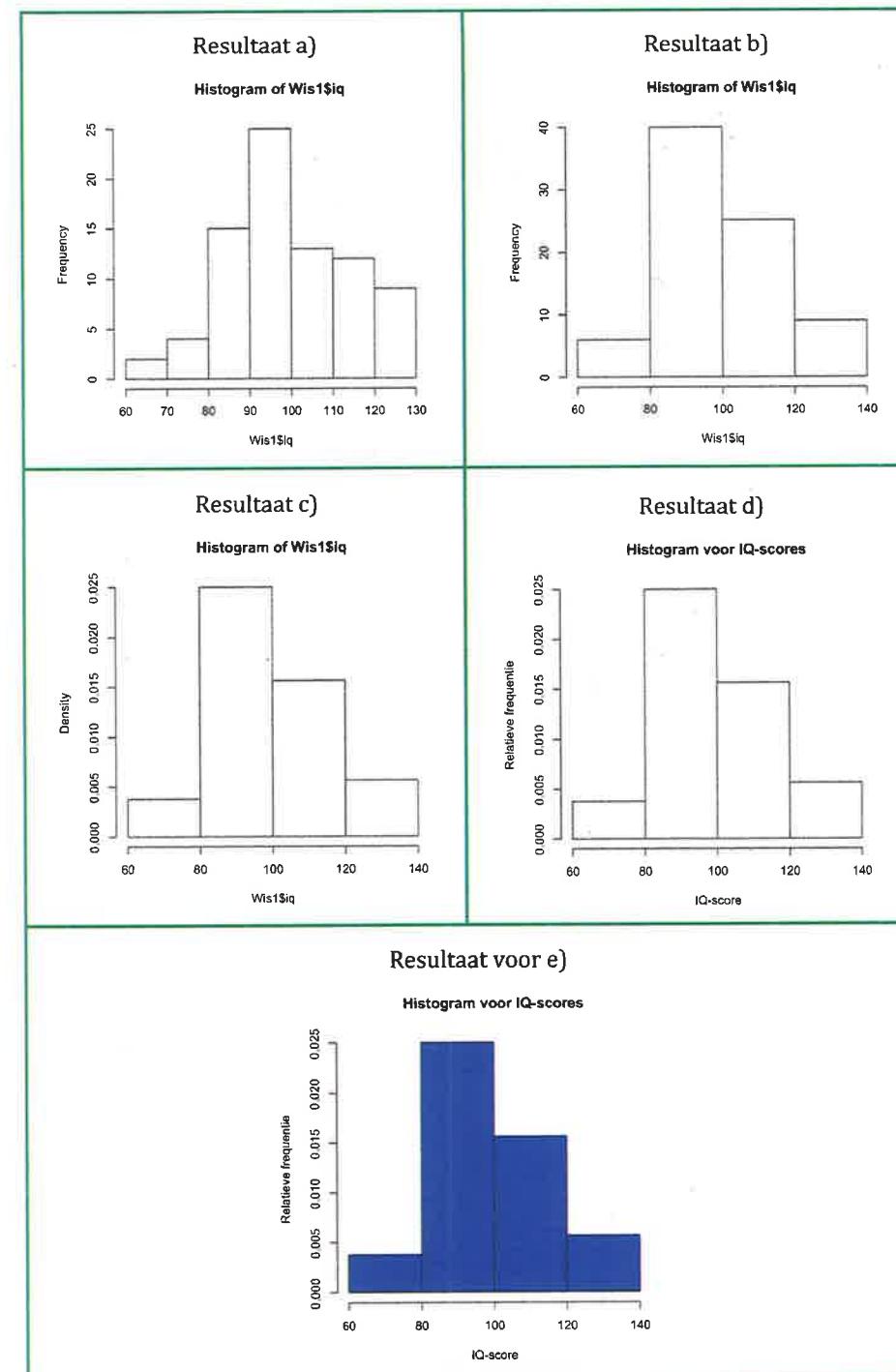
41	119.01	1	1.25	68	85.00
42	119.98	3	3.75	71	88.75
43	120.96	1	1.25	72	90.00
44	121.93	1	1.25	73	91.25
45	123.87	3	3.75	76	95.00
46	124.84	2	2.50	78	97.50
47	125.81	1	1.25	79	98.75
48	128.73	1	1.25	80	100.00

- b) In totaal zijn er 48 verschillende IQ-scores terug te vinden in dit bestand.
- c) Er zijn 2 respondenten die een IQ-score 100.56 hebben
- d) De waarde 119,98 komt voor op de 42e rij in bovenstaande tabel en komt overeen met een percentage van 3,75%. 3,75% van de respondenten behaalde dus een IQ-score van 119,98. Merk op dat de relatieve frequentie van de waarde 119,98 gelijk is aan $f_{42} = 0,0375$.
- e) De waarde 90 staat niet in de frequentietabel. Daarom kijken we naar de eerste waarde daarvoor (89,87). Door vervolgens naar de relatieve cumulatieve frequentie voor die waarde te kijken, kunnen we het antwoord weten: 26,25%.

Respons 5.4.3

De volgende commando's leveren je de verschillende versies van de histogrammen op:

- a) > hist(Wis1\$iq)
- b) > hist(Wis1\$Iq, nclass=4)
- c) > hist(Wis1\$Iq, nclass=4, freq=FALSE)
- d) > hist(Wis1\$Iq, nclass=4, freq=FALSE, xlab='IQ-score', +
y lab='Relatieve frequentie', main='Histogram voor IQ-scores')
- e) > hist(Wis1\$Iq, nclass=4, freq=FALSE, xlab='IQ-score', +
y lab='Relatieve frequentie', main='Histogram voor +
IQ-scores', col='blue')



Respons 5.5.3

a) Om een staafdiagram te maken met absolute frequenties gaan we als volgt te werk:

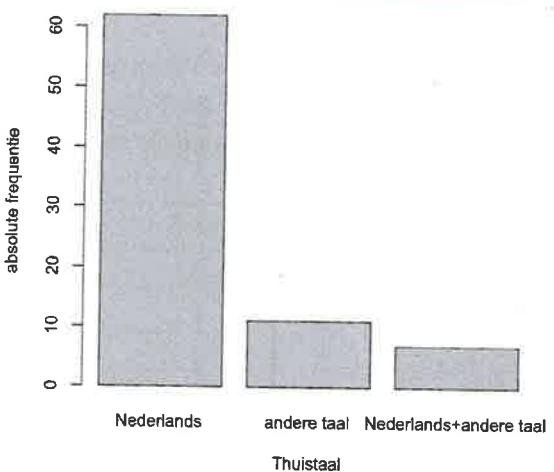
Eerste stap: een object aanmaken met daarin het resultaat van `table()`.

```
> FreqT<-table(Wis1$Thuistaal)
```

Tweede stap: het staafdiagram aanmaken.

```
> barplot(sort(FreqT, decreasing=TRUE), xlab='Thuistaal',
ylab='Absolute frequentie')
```

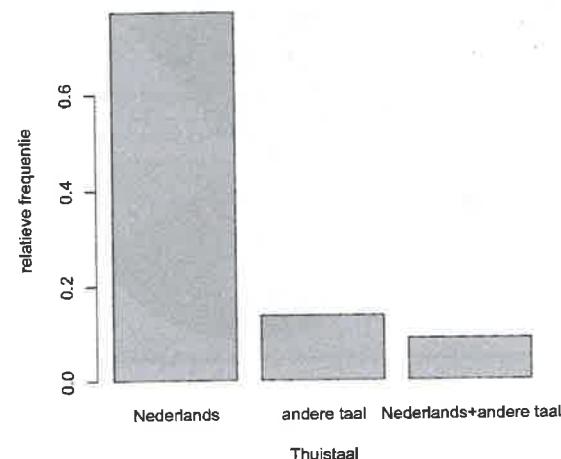
Hieronder het resultaat:



b) Om een vergelijkbare figuur aan te maken, maar met relatieve frequenties maken we bovendien gebruik van het commando `prop.table()`.

```
> barplot(prop.table(sort(FreqT, decreasing=TRUE)), xlab='Thuistaal',
ylab='relatieve frequentie')
```

Hieronder het resultaat:



Respons 5.5.6

Om dit te bekomen dien je de volgende commando's te hanteren:

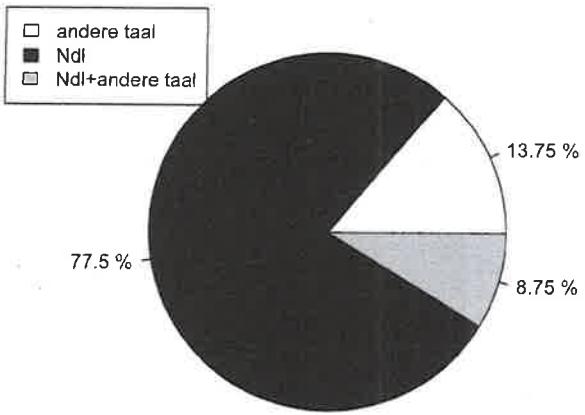
```
# Aanmaken van een object met naam Freq2 waarin het resultaat van de
# table() functie staat
> Freq2 <- table(Wis1$Thuistaal)

# Een object met de naam kleurtjes aanmaken waarvan we straks gebruik
# kunnen maken bij het opstellen van de grafiek
> kleurtjes <- c('white', 'grey20', 'grey80')

# Een object aanmaken waarin we de relatieve frequenties bewaren om
# te hanteren als labels bij de taartpunten
> Eigenlabels <- prop.table(Freq2)*100
> Eigenlabels <- round(Eigenlabels, 2)
> Eigenlabels <- paste(eigenlabels, "%", sep=" ")

# Het aanmaken van de grafiek
> pie(Freq2, col=kleurtjes, labels=Eigenlabels)

# Een legende toevoegen (laat de grafiek ondertussen open staan), die
# horizontaal op waarde -1.45 gepositioneerd is en verticaal op 1.
# Met beide cijfers moet je een beetje spelen om eens het gevolg
# ervan te zien op de plaats waar de legende komt te staan.
> legend(-1.45, 1, c('andere taal', 'Ndl', 'Ndl+andere taal'),
fill=kleurtjes)
```

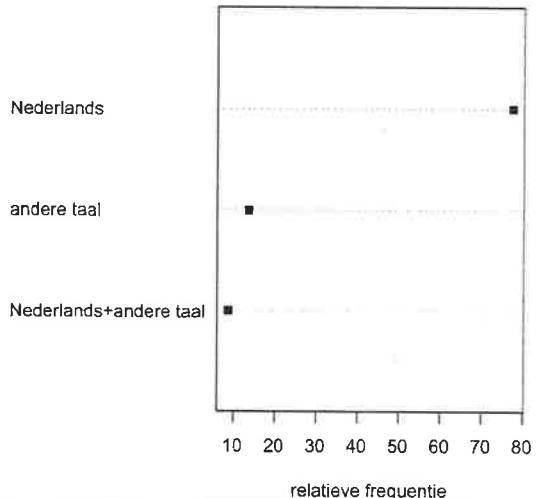


Respons 5.5.9

Om dit te bekomen dien je de volgende commando's te hanteren:

```
> Freq2 <- table(Wis1$Thuisstaal)
> dotchart(sort(prop.table(Freq2)*100, decreasing=FALSE),
xlab='relatieve frequentie', pch=15)
```

Let in het commando op het argument `decreasing=FALSE` binnen het `sort()` commando. Dit zorgt ervoor dat de categorieën in oplopende volgorde worden sorteerd. Hieronder het resultaat dat je zou moeten bekomen.



Gehanteerde functies

Functie	Doelstelling	Bron
<code>barplot()</code>	Maakt een staafdiagram voor een categorische variabele. LET OP: tussen haakjes het resultaat van een <code>table()</code> schrijven. vb: <code>barplot(table(Wis1\$Score))</code>	R basispakket
<code>dotchart()</code>	Maakt een puntdiagram voor een categorische variabele. LET OP: tussen haakjes het resultaat van een <code>table()</code> schrijven. vb: <code>dotchart(table(Wis1\$Score))</code>	R basispakket
<code>fregtabel()</code>	Geeft een volledige frequentietabel voor een variabele vb: <code>fregtabel(Wis1\$lq)</code>	OLP Functies.R
<code>hist()</code>	Maakt een histogram voor een variabele	R basispakket
<code>paste()</code>	Maakt het mogelijk om stukken tekst achter elkaar te zetten.	R basispakket
<code>pie()</code>	Maakt een taartdiagram voor een categorische variabele. LET OP: tussen haakjes het resultaat van een <code>table()</code> schrijven. vb: <code>pie(table(Wis1\$Score))</code>	R basispakket
<code>plot()</code>	Maakt een gepasteerde grafiek van een variabele. Voor numerieke variabelen wordt een scatterplot (puntgrafiek) gemaakt. Voor categorische variabelen ("factors") wordt een staafdiagram gemaakt. vb: <code>plot(Wis1\$lqcategor)</code>	R basispakket
<code>prop.table()</code>	Geeft de relatieve frequentie, uitgedrukt in een proportie (tussen 0 en 1) van de meetwaarden van een variabele. LET OP: tussen haakjes het resultaat van een <code>table()</code> schrijven. vb: <code>prop.table(table(Wis1\$Score))</code>	R basispakket
<code>round()</code>	Maakt het mogelijk om waarden af te ronden tot op een gekozen aantal decimalen. vb: <code>round(Wis1\$Score,2)</code>	R basispakket
<code>sort()</code>	Sorteert de elementen uit een object dat het resultaat is van een <code>table()</code> commando. Werd geïntroduceerd om de volgorde van categorieën in staafdiagrammen te laten afhangen van de frequentie. vb: <code>MijnTabel<-table(Wis1\$Score)</code> <code>barplot(sort(MijnTabel))</code> Sorteren gebeurt standaard van klein naar groot. Ordenen van groot naar klein kan door het argument 'decreasing = TRUE' te gebruiken.	R basispakket
<code>table()</code>	Geeft de absolute frequenties van de meetwaarden van de variabele	R basispakket

HOOFDSTUK 6

Parameters van ligging en spreiding

DOELSTELLINGEN:

Na dit hoofdstuk:

- kan je de verschillende kengetallen van zowel ligging (modus, mediaan, rekenkundig gemiddelde, kwantielen) als van spreiding (reikwijdte, interkwartielafstand, gemiddelde absolute afwijking, variantie en standaardafwijking) uitleggen;
- weet je het meest gepaste kengetal voor te stellen voor de verschillende soorten variabelen;
- kan je deze verschillende kengetallen via R zowel stapsgewijs als aan de hand van een specifieke functie berekenen;
- ben je in staat om een boxplot te produceren en te interpreteren.

NODIGE FILES:

Pirls1.RData

een file met daarin een aantal variabelen over 137 Vlaamse scholen die deelnamen aan een internationaal vergelijkend onderzoek.

OLP Functies.R

een file met daarin aangepaste functies die bij dit OLP horen.



In de beschrijvende statistiek wordt gebruik gemaakt van een aantal parameters, ook soms kengetallen genoemd, die iets zeggen over de ligging of de locatie van de waarnemingen en de mate dat er verschillen zijn tussen de waarnemingen.

Een **kengetal** is een cijfer dat op zich een samenvatting geeft van de verschillende waarnemingen. Met andere woorden, aan de hand van één cijfer vat je bepaalde eigenschappen van de variabelen samen. In dit hoofdstuk bespreken we achtereenvolgens parameters van ligging en parameters van spreiding. Daarna richten we onze aandacht op de grafische representatie van zowel ligging als spreiding aan de hand van de boxplot.

6.1. Parameters van ligging

6.1.1

We starten dit hoofdstuk met de kengetallen die we kunnen hanteren om de plaats van de waarnemingen te beschrijven.



Een alternatieve benaming van deze parameters is: centrale tendentiematen of centrummaten.

Het gebruik van het meervoud duidt er reeds op dat er meer dan één manier is om de ligging van de waarnemingen te beschrijven. Er zijn met andere woorden meerdere kengetallen die iets zeggen over de ligging.

De **keuze van welke maat** je hanteert om de ligging te beschrijven is **afhankelijk van het meetniveau** van de variabele. In wat volgt bespreken we eerst de centrummaten voor nominale variabelen, gevolgd door de centrummaten voor ordinale variabelen en voor interval- en ratiovariabelen.

Hierbij geldt de regel dat je de centrummaten die je hanteert bij een variabele van lager meetniveau mag gebruiken bij een variabele van een hoger meetniveau, maar niet omgekeerd. Je mag dus een centrummaat voor ordinale variabelen wel hanteren voor intervalvariabelen, maar niet omgekeerd.

6.1.2

Om de ligging van een **nominale variabele** te beschrijven wordt gebruik gemaakt van de modus.



De **modus** is niet meer of niet minder dan de categorie die het meeste voorkomt in de waarnemingen.

In de onderstaande tabel geven we de frequentieverdeling van de Vlaamse leerlingen uit de PISA 2003 databank naar de variabele 'arbeidsmarktpositie van hun moeder'.

Tabel 6.1: Frequentieverdeling van de Vlaamse leerlingen uit de PISA 2003 databank naar de arbeidsmarktpositie van hun moeder

Arbeidsmarktpositie moeder	Frequentie (ni)	Percentage (fi)
Werkt voltijs	2123	43,8
Werkt deeltijs	1255	25,9
Werkzoekende	158	3,3
Andere	1314	27,1
Totaal	4850	100,0

De modus voor deze variabele is: werkt voltijs. Dit is de categorie die het meeste voorkomt in de gegevens. Met andere woorden, als ze je zouden vragen om te raden naar de arbeidsmarktpositie van de moeder van een willekeurige Vlaamse leerling, dan kan je best het antwoord 'voltijs werken' geven. Met dat antwoord heb je de grootste kans dat je het juist hebt.

Het is niet ondenkbaar dat er voor een bepaalde variabele meerdere **modi** zijn: twee of meer categorieën komen het meest voor. In dat geval noemen we de verdeling voor deze variabele **multimodaal**. Indien er slechts één modus is, noemen we de verdeling **unimodaal**.

6.1.3



Het databestand Pirls1.RData bevat een reeks variabelen over 137 Vlaamse scholen die deelnamen aan een internationale studie over het leesonderwijs (Pirls-studie).

Eén van die variabelen is de 'Stedelijkheid' (Pirls1\$Stedelijkheid). Deze variabele behelst drie categorieën: stad, randstad en landelijk.

Ga via R na wat de modus is.

Tip: pas toe wat je geleerd hebt in het hoofdstuk over de frequentieverdeling (hoofdstuk 4).

6.1.4



Willen we de ligging van een **ordinale variabele** beschrijven dan kunnen we naast de modus de mediaan hanteren.

De **mediaan** is de waarde voor de variabele waarop we de geordende antwoorden van de respondenten in twee gelijke groepen kunnen verdelen.

Dit is bijgevolg de meetwaarde waarvoor we evenveel respondenten kunnen vinden die hoger scoren dan deze meetwaarde als dat we respondenten kunnen vinden die lager scoren.

Het is eigen aan ordinale variabelen dat we de waarnemingen kunnen rangschikken in een oplopende volgorde. Als we dit doen is de mediaan de meetwaarde van de middelste waarneming.

Formeel drukken we de mediaan uit als: **Me**.

$$Me = \text{de } \left(\frac{n+1}{2} \right)^{\text{de}} \text{ waarde}$$

Een voorbeeld kan dit verduidelijken. Stel dat we aan negen respondenten hebben gevraagd om aan te geven in welke mate ze het eens zijn met een bepaalde stelling. Ze konden voor het antwoorden de volgende vijfpuntschaal gebruiken: 1=helemaal oneens; 2=oneens; 3=noch oneens/noch eens; 4=eens; 5=helemaal eens.

We stellen daarbij het volgende antwoordpatroon vast:

2	3	4	2	5	1	2	2	3
---	---	---	---	---	---	---	---	---

We rangschikken deze waarnemingen in oplopende volgorde:

1	2	2	2	2	3	3	4	5
---	---	---	---	---	---	---	---	---

Passen we de formule $(n+1)/2$ toe: $(9+1)/2=5$. De vijfde waarneming is een '2'. In dit voorbeeld zouden we de mediaan dus gelijkstellen aan de categorie twee 'oneens'.

De berekening is niet altijd zo makkelijk. Stel dat we dezelfde reeks hebben als in het bovenstaande voorbeeld, maar met één waarneming minder (een twee die wegvalt). Als je de waarnemingen vervolgens rangschikt krijg je het volgende patroon:

1	2	2	2	3	3	4	5
---	---	---	---	---	---	---	---

We passen opnieuw de formule $(n+1)/2$ toe: $(8+1)/2=4,5$. De mediaan is dus de $4,5^{\text{de}}$ waarde. Deze waarde hebben we echter niet waargenomen.

In dat geval wordt de mediaan vastgelegd op de waarde die in het midden ligt tussen de vierde en de vijfde waarde. Hier dus het getal dat midden tussen twee en drie ligt. De mediaan is in dit voorbeeld bijgevolg 2,5. De **mediaan kan dus een waarde aannemen die niet rechtstreeks waargenomen is.**

Bijgevolg kan de mediaan een fictieve waarde aannemen. Dit is zo bij discrete variabelen. Dit zijn variabelen waarbij tussen twee willekeurige meetwaarden niet noodzakelijk een andere meetwaarde ligt. Een voorbeeld: stel dat de onderstaande acht waarnemingen het aantal jongens in een klas weergeven:

11	12	12	13	14	15	17	20
----	----	----	----	----	----	----	----

De mediaan voor deze variabele bedraagt 13,5. Dit is een fictieve waarde want het is onmogelijk dat er 13,5 jongens in een klas zitten.

De mediaan is niet altijd even informatief. Stel dat onze negen respondenten het volgende antwoordpatroon hadden vertoond:

1	2	2	2	2	2	2	2	2
---	---	---	---	---	---	---	---	---

Bereken je enkel de mediaan en kijk je niet verder naar de gegevens dan zou dit kunnen leiden tot een volgende conclusie: de helft van de respondenten behaalt een score hoger dan twee. Echter acht van de negen respondenten antwoordt hetzelfde, categorie twee.

6.1.5



Het databestand Pirls1.RData bevat een variabele die het antwoord bevat van de 137 directies op de vraag "Hoe vaak komen leerkrachten formeel samen om materialen uit te wisselen of samen te ontwikkelen?". Deze variabele heet "Samen". Om daarop een antwoord te formuleren konden de directies kiezen tussen de volgende categorieën:

- 1 = minstens 2-3 keer per week;
- 2 = één keer per week;
- 3 = één keer per maand;
- 4 = minder dan één keer per maand;
- 5 = nooit.

Maak aan de hand van de functie `freqtabel()` (zie vorig hoofdstuk) uit het bestand OLP Functies.R een frequentietabel aan. Kan je aan de hand van die tabel de mediaan afleiden?

6.1.6



In R kunnen we ook gebruik maken van een eenvoudige functie om de mediaan te achterhalen: `median()`. Echter, een belangrijk aandachtspunt daarbij is dat deze functie in R enkel toegepast kan worden op numerieke variabelen. Als we, met andere woorden, te maken hebben met een categorische variabele (van ordinale niveau) (zoals in 6.1.5) dan moeten we er eerst voor zorgen dat R deze variabele beschouwt als numerieke variabele. Daartoe kunnen we eenvoudig gebruik maken van de functie `as.numeric()`. Deze functie maakt van een categorische variabele een numerieke variabele.

Om nu de mediaan op te vragen van bijvoorbeeld de variabele "Samen" (zie 6.1.5) kunnen we beide commando's combineren. Dit wordt dan:

```
> median(as.numeric(Pirls1$Samen))
```

Als je dit commando ingeeft in R krijg je het volgende resultaat:

```
[1] NA
```

Dit is uiteraard niet wat we beoogden. De reden dat we dit uitkomen, ligt in het feit dat er voor de variabele "Samen" ontbrekende waarden (NA's) zijn. Sommige directies bleven ons het antwoord schuldig op die vraag. We moeten in onze functie aangeven dat we willen dat R deze NA's verwijdert. Dit doen we door het argument `na.rm=TRUE` toe te voegen aan het commando. Hierbij staat rm voor remove. We geven met andere woorden mee aan R dat de NA's verwijderd mogen worden:

```
> median(as.numeric(Pirls1$Samen), na.rm=TRUE)
[1] 3
```

De mediaanwaarde is dus drie, net zoals we dat zelf ook concludeerden uit de frequentietabel van 6.1.5.

6.1.7



Bij **intervalvariabelen** kunnen zowel de mediaan als de modus gehanteerd worden als centrummaat. Daarnaast is er een centrummaat die we alleen voor intervalvariabelen gebruiken. Dit is voor de meeste mensen een bekende maat: **het gemiddelde**. De waarde die we krijgen door alle waargenomen meetwaarden voor een variabele op te tellen en te delen door het aantal waarnemingen. Wat de meeste mensen verstaan onder de term 'het gemiddelde' is wat we meer precies het **rekenkundig gemiddelde** noemen.

Formeel kan het rekenkundig gemiddelde als volgt geschreven worden:

$$\mu = \frac{\sum_{i=1}^n x_i}{n}$$

Deze formule lezen we als volgt. x_i staat voor de i -de waarneming voor de variabele X . In totaal zijn er n waarnemingen. De teller lezen we dus als de som van alle waarnemingen van de variabele X . De noemer bevat het aantal waarnemingen.

Net als de mediaan kan het rekenkundig gemiddelde een **meetwaarde aannemen die we niet werkelijk waargenomen hebben**. Het rekenkundig gemiddelde voor een discrete variabele kan dus een fictieve waarde hebben. Bijvoorbeeld wanneer het gemiddelde aantal jongens per klas in alle Vlaamse scholen 11,4 bedraagt.

Een belangrijke eigenschap van de mediaan is dat deze centrummaat **NIET gevoelig is voor uitbijters** (=extreme meetwaarden). Dit maakt dat deze maat ook vaak gehanteerd wordt voor intervalvariabelen. Het gemiddelde is **WEL gevoelig aan uitbijters**. Extreme meetwaarden voor een variabele kunnen het gemiddelde sterk omhoog of omlaag trekken.

6.1.8



Het rekenkundig gemiddelde kunnen we vrij makkelijk in R uitrekenen door gebruik te maken van twee functies: één voor de teller en één voor de noemer. In de teller staat de som van alle observaties. Om de som te nemen van alle waarden maken we gebruik van de `sum()` functie. Bij deze functie is het wederom van belang om aan te geven dat R de NA's (missende waarden) moet negeren. Dit doen we door het extra argument `na.rm=TRUE` toe te voegen:

```
> sum(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 41605
```

Als we dus alle geobserveerde schoolgroottes met elkaar optellen overeen alle geobserveerde scholen, dan resulteert dit in een waarde van 41605 leerlingen.

Vervolgens willen we dit delen door het aantal valide waarnemingen. Met valide waarnemingen bedoelen we hier het aantal observatie-eenheden (scholen) waarvan we een effectieve waarde hebben (geen NA's). Er zijn verschillende wegen om dit getal te achterhalen. Een relatief eenvoudige

wijze is gebruik maken van de functie `length()`. Deze functie geeft de lengte van de variabele. Dit is het aantal rijen in een variabele. Let op, daarbij worden de NA's ook meegeteld. Toegepast voor de variabele "schoolgrootte" geeft dit:

```
> length(Pirls1$Schoolgrootte)
[1] 137
```

Er zijn dus 137 rijen in onze dataset. Echter, het is mogelijk dat we van niet alle 137 scholen de schoolgrootte kennen. Er zijn dus mogelijk scholen met een NA voor deze variabele. Die willen we uiteraard niet meenemen in de berekening van het gemiddelde. Daarom moeten we wederom iets doen met de ontbrekende waarden (NA's). Daartoe kunnen we gebruik maken van een specifieke functie die alle NA's verwijdert terwijl we gebruik maken van de `length()` functie: `na.omit()`. We combineren beide functies en dat geeft:

```
> length(na.omit(Pirls1$Schoolgrootte))
[1] 128
```

We hebben nu beide ingrediënten om het rekenkundig gemiddelde te berekenen. Door beide onderdelen in één commando te combineren krijgen we het gemiddelde:

```
> sum(Pirls1$Schoolgrootte, na.rm=TRUE) /
length(na.omit(Pirls1$Schoolgrootte))
[1] 325.0391
```

De gemiddelde schoolgrootte van de school bedraagt dus 325 leerlingen. Nemen we een willekeurige school uit het bestand en willen we daarvoor een gok wagen wat de schoolgrootte is van die school, dan is de beste gok het gemiddelde: 325.

6.1.9



Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het 4^{de} leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

- Voor hoeveel scholen kennen we het aantal pc's niet?
- Hoeveel pc's zijn er aanwezig in alle geobserveerde scholen samen?
- Wat is het gemiddeld aantal pc's in onze steekproef?

-  **6.1.10** In 6.1.8 toonden we hoe het rekenkundig gemiddelde uitgerekend kan worden in R. We kunnen echter makkelijker, net zoals bij de mediaan, het rekenkundig gemiddelde opvragen via één functie: `mean()`. Deze functie werkt hetzelfde als de `median()` functie. Indien er NA's zijn, moeten we gebruik maken van het argument `na.rm=TRUE`. Hieronder wordt dit toegepast voor schoolgrootte:

```
> mean(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 325.0391
```

De functies die we tot hertoe hanteerden, zowel via de stapsgewijze berekening als via het ene commando (bv. `mean()` of `median()`), stellen ons in staat om het resultaat ook effectief weg te schrijven naar een ander object. Zo zouden we bijvoorbeeld aan de hand van het volgende commando het gemiddelde van schoolgrootte kunnen wegschrijven in een afzonderlijk object met de naam "Gem_schoolgrootte".

```
> Gem_schoolgrootte <- mean(Pirls1$Schoolgrootte, na.rm=TRUE)
```

Dit object kunnen we dan op elk moment opnieuw oproepen of hanteren in andere bewerkingen. Later zullen we dit meermalen toepassen.

Indien we echter niet geïnteresseerd zijn in het wegschrijven van de mediaan of het gemiddelde dan kunnen we voor kwantitatieve variabelen in één beweging zowel het gemiddelde als de mediaan opvragen aan de hand van de `summary()` functie:

```
> summary(Pirls1$Schoolgrootte)

Min.   1st Qu.    Median      Mean   3rd Qu.      Max.     NA's
26.0      233.0     308.5     325.0     405.5     807.0      9.0
```

Hieruit leren we dat de mediaan 308.5 is en het gemiddelde 325. We lezen ook meteen af dat er 9 scholen zijn waarvan we de schoolgrootte niet kennen. Daarnaast staan er nog een aantal andere samenvattende gegevens in deze output. Die zullen we later in dit hoofdstuk bespreken.

-  **6.1.11** Naast het bekende rekenkundig gemiddelde bestaan er ook twee andere gemiddelden: het harmonisch en het geometrisch gemiddelde. Deze gemiddelden worden minder vaak toegepast binnen de sociale wetenschappen. Voor de volledigheid bespreken we deze twee andere gemiddelden eveneens.

Het **harmonisch gemiddelde** van een aantal waarnemingen wordt verkregen door de inverse van de meetwaarden bij elkaar op te tellen en vervolgens het aantal observaties te delen door dit totaal.

De volgende formule drukt dit formeel uit:

$$\mu_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Het harmonisch gemiddelde wordt gebruikt indien de meetwaarden uitgedrukt zijn in relatie tot een andere grootheid. Het klassieke voorbeeld zijn snelheden. De gemiddelde snelheid van tweeritten kan worden bepaald met het harmonisch gemiddelde, gegeven dat deritten even lang zijn qua afstand. Als de heenreis wordt gereden aan 100 km/u en de terugreis aan 120 km/u, dan is de gemiddelde snelheid van de totale rit volgens het harmonisch gemiddelde 109,09 km/u i.p.v. 110 km/u wat je zou krijgen bij het berekenen van het rekenkundig gemiddelde.

Dit wordt duidelijk indien je dit in een tabel naast elkaar zet:

Tabel 6.2: Uitwerking van het harmonisch gemiddelde voor de snelheden

	tijd	snelheid	afstand
heenreis	30 min	100 km/u	50 km
terugreis	25 min	120 km/u	50 km
totaal	55 min		100 km
harmonisch gemiddelde	$(100/55)*60 = 109,09 \text{ km/u}$		

Beide reizen samen vormt een reis van 100 km die 55 minuten duurde. Dit brengt de gemiddelde snelheid voor deze volledige reis op 109,09 km/u. Het rekenkundig gemiddelde van beide snelheden (110 km/u) had tot een verkeerde conclusie geleid.

Het **geometrisch gemiddelde** van n meetwaarden is gelijk aan de n -de machtswortel uit het product van die uitkomsten. In formulevorm ziet dit er als volgt uit:

$$\mu_g = \sqrt[n]{\prod_{i=1}^n x_i} = \sqrt[n]{x_1 x_2 x_3 \dots x_n}$$

Het geometrisch gemiddelde wordt gebruikt indien we het gemiddelde willen berekenen van verschillende relatieve meetwaarden (percentages). Een voorbeeld: stel dat je van een school voor vijf jaren hebt vastgesteld met hoeveel procent de schoolbevolking is toegenomen of afgenaomen. Om de gemiddelde toename per jaar over deze vijf jaar te berekenen dien je het geometrisch gemiddelde te berekenen. Het volgende voorbeeld met een schoolgrootte van 200 leerlingen in jaar nul verduidelijkt dit.

Tabel 5.3: Uitwerking van het geometrisch gemiddelde voor de schoolgroottes

Jaar	Toename in percenten	Relatieve toename	Schoolgrootte
0			200
1	13%	1,13	226 (=200*1,13)
2	15%	1,15	260 (=226*1,15)
3	12%	1,12	290 (=260*1,12)
4	-5%	0,95	275 (=290*0,95)
5	-13%	0,87	239 (=275*0,87)

Het geometrisch gemiddelde berekenen we als volgt:

$$\mu_g = \sqrt[5]{1,13 * 1,15 * 1,12 * 0,95 * 0,87} = 1,04$$

Het geometrisch gemiddelde spiegelt zich rond de waarde één. Hier is het geometrisch gemiddelde 1,04. Dit kunnen we interpreteren als: de schoolgrootte nam de afgelopen vijf schooljaren gemiddeld met 4% per jaar toe. In totaal nam de schoolgrootte bijna 20% toe over deze vijf schooljaren ($239/200=1,195$ of 20% toename). Hadden we een geometrisch gemiddelde van 0,95 behaald, dan wou dit zeggen dat de schoolgrootte de afgelopen vijf jaar met gemiddeld 5% per jaar was afgenaomen.

Het geometrisch gemiddelde vormt vaak een alternatief voor het rekenkundig gemiddelde indien een variabele een rechtsscheve verdeling heeft. Hierover later meer. In feite komt het erop neer dat in zo'n geval het geometrisch gemiddelde duidelijk minder gevoelig is voor extreme hoge meetwaarden voor de variabele. Bijgevolg geeft het geometrisch gemiddelde dan een 'zuiverder' beeld van de centraliteit van zo'n variabele.

6.1.12

i Tot hiertoe hebben we enkel parameters van ligging besproken die zeer specifiek inzoomen op de centrale waarneming. Naast de modus, de mediaan en het gemiddelde kunnen we voor **ordinale variabelen** ook gebruik maken van andere kengetallen die iets zeggen over de ligging van onze

observaties. Deze kengetallen zijn: kwartieLEN, decielEN en percentielEN. Meer generiek worden deze kengetallen als groep "kwantieLEN" genoemd.

KwartieLEN verdelen geordende observaties in 4 gelijke delen. KwartieLEN worden aangeduid met K1, K2, K3 (of Q1, Q2, Q3).

K1 is die waarde waarvan we kunnen zeggen dat 25% van de observaties kleiner dan of gelijk is aan die meetwaarde. K2 is in feite net hetzelfde als de mediaan: 50% van de observaties behalen deze of een lagere meetwaarde. K3 ten slotte is die meetwaarde waar we voor vaststellen dat 25% van de observaties hoger scoort dan die meetwaarde.

De berekening van kwartieLEN kan meer formeel worden gevatt door de functies:

$$K1 = \frac{n+1}{4} \text{ de meetwaarde;}$$

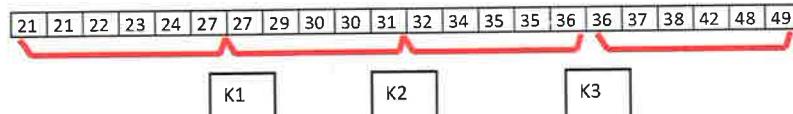
$$K2 = \frac{2(n+1)}{4} \text{ de meetwaarde;}$$

$$K3 = \frac{3(n+1)}{4} \text{ de meetwaarde.}$$

Aan de hand van een voorbeeld wordt dit duidelijk. Stel we hebben 22 waarnemingen van leeftijden:

22	21	30	36	30	35	27	27	29	42	21
31	38	37	23	35	32	48	36	24	49	34

Om de kwartieLEN, decielEN of percentielEN te berekenen, dienen deze observaties te worden gerangschikt van laag naar hoog.



De kwartieLEN zijn hier respectievelijk:

$$K1 = 27$$

$$K2 = 31,5$$

$$K3 = 36$$

De meetwaarden die geordende waarnemingen in 10 gelijke delen indelen worden **deciliën** genoemd (van D1 tot en met D9), terwijl de waarden die ordinale gegevens in 100 gelijke delen indelen de **percentielen** worden genoemd (van P1 tot P99)..

Ook hier stemmen het 5^{de} deciel of het 50^{ste} percentiel overeen met de mediaan.

6.1.13



In R is er een generieke functie `quantile()` die ons in staat stelt om allerlei kwantielen op te vragen. Belangrijk bij deze functie, net als bij de `median()` en `mean()` functie, is het aangeven dat R de NA's moet verwijderen. Daartoe hanteren we het argument `na.rm=TRUE`.

Standaard geeft dit commando de drie kwartilen (en de minimum en maximum geobserveerde meetwaarde). Een voorbeeld maakt dit duidelijk. Hernemen we de Pirls1 dataset en meer specifiek de variabele schoolgrootte dan zien het standaardcommando en het bijbehorende resultaat er als volgt uit:

```
> quantile(Pirls1$Schoolgrootte, na.rm=TRUE)
0%    25%    50%    75%   100%
26.0   233.0  308.5  405.5  807.0
```

K1 bedraagt dus 233. Eén op de vier scholen telt met andere woorden 233 leerlingen of minder. K3 bedraagt 405,5: één op de vier scholen telt meer dan 405,5 leerlingen.

Dezelfde functie kunnen we ook hanteren om deciliën of percentielen op te vragen. Daartoe voegen we een extra argument toe onder de vorm van `c()`. Tussen de haakjes geef je de percentages weer waarvoor je meetwaarden wil. Stel bijvoorbeeld dat we enkel en alleen het eerste deciel willen weten (=10% scoort hetzelfde of lager) dan vertaalt dit zich in het opnemen van het volgende argument `c(.10)`. We passen het hieronder toe:

```
> quantile(Pirls1$Schoolgrootte, c(.10), na.rm=TRUE)
10%
155.1
```

10% van de scholen telt 155,1 leerlingen of minder.

Willen we naast het 1^{ste} deciel ook het negende deciel weten en eveneens de mediaan dan ziet dit extra argement er zo uit c (.10, .50, .90).

```
> quantile(Pirls1$Schoolgrootte, c(.10, .50,.90), na.rm=TRUE)
10%    50%    90%
155.1  308.5  488.6
```

6.1.14



Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het vierde leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

a) D1 = ... ?

b) K3 = ... ?

c) D9 = ... ?

d) K2 = ... ?

e) Wat is het maximum aantal pc's in de 25% laagst scorende scholen voor deze variabele?

f) Wat is het maximum aantal pc's in de 32% laagst scorende scholen voor deze variabele?

6.2. Parameters van spreiding

6.2.1



We worden overspoeld met maatschappelijk relevante cijfers. Daarbij worden we vaak getrakteerd op uitspraken over de kenmerken van de "modale burger", de "gemiddelde jongere", zonder te weten over wie het precies gaat, of hoe het kenmerk rond dat gemiddelde is verdeeld. Dit levert vaak zeer beperkte informatie op en dit wordt statistici vaak kwalijk genomen. Zo lanceerde Godfried Bomans het volgende gezegde:

"Een statisticus waadde vol vertrouwen door een rivier die gemiddeld één meter diep was. Hij verdronk."

Of een andere, van N. Locke:

"Een statisticus is iemand die, als hij met zijn hoofd in een brandende kachel en met zijn voeten in een emmer ijs staat, verklaart: 'Gemiddeld genomen, voel ik me wel lekker'."

Naast de ligging is het ook vaak van belang een beeld te krijgen van de spreiding. Daarbij doen we op het in kaart brengen van de grootte van de verschillen in de geobserveerde meetwaarden. Om bij het beeld van Godfried Bomans te blijven: hoeveel verschil is er tussen de verschillende plaatsen in de rivier qua diepte.

In dit OLP zullen we ons beperken tot het bespreken van kengetallen van spreiding die we kunnen toepassen op variabelen van ordinaal meetniveau of hoger.

6.2.2



Om een eerste idee te krijgen van al de mogelijke verschillen in een reeks waarnemingen kan men de **variatiebreedte (V)** van een kenmerk in kaart brengen. Deze variatiebreedte, of “reikwijdte” (in het Engels “range”) is het verschil tussen de grootste en de kleinste waargenomen meetwaarde. De reikwijdte wordt dus berekend door het minimum en het maximum te vergelijken, of formeel

$$V = \max(X) - \min(X)$$

Hernemen we het voorbeeld van de 22 leeftijden uit 6.1.12 dan wordt de variatiebreedte gegeven door

$$V = 49 - 21 = 28$$

Met andere woorden, de oudste en de jongste respondent verschillen 28 jaar van elkaar.

De variatiebreedte geeft al een eerste indicatie van de spreiding van een variabele. Vertaald naar de rivier van Bomans: de ietwat voorzichtige en geïnformeerde statisticus zal wel opletten om een rivier over te steken die gemiddeld 1m diep is en een variatiebreedte van meer dan 2m heeft. Dat wil immers zeggen dat de rivier ergens minstens 2m diep is.

De reikwijdte kan je uiteraard enkel zinvol toepassen bij variabelen van interval- of rationale niveau. Immers, wat zou de betekenis kunnen zijn van de reikwijdte van een variabele die gemeten is aan de hand van een vijf-puntenschaal gaande van één (helemaal oneens) tot vijf (helemaal eens)? Verschillen zijn niet zinvol bij een ordinale variabele.

6.2.3



In R kunnen we van drie functies gebruik maken om de reikwijdte na te gaan: `min()` geeft de minimumscore; `max()` geeft de maximumscore; `range()` geeft in één beweging zowel de minimum- als maximumscore.

We passen dit hieronder toe (let daarbij op het argument `na.rm=TRUE`):

```
> min(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 26
> max(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 807
> range(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 26 807
```

We kunnen ook in een beweging de reikwijdte uitrekenen als volgt:

```
> max(Pirls1$Schoolgrootte, na.rm=TRUE) -
  min(Pirls1$Schoolgrootte, na.rm=TRUE)
[1] 781
```

Het bestand “OLP functies.R” bevat de functie: `reikwijdte()`. Aan de hand van deze functie kan je in één commando de reikwijdte opvragen.

```
> reikwijdte(Pirls1$Schoolgrootte)
[1] 781
```

6.2.4



Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het 4^{de} leerjaar basisonderwijs. Deze variabele heeft de naam “Pc”.

- a) Wat is het minimum en het maximum aantal pc's beschikbaar op de geobserveerde scholen?
- b) Wat is de reikwijdte van deze variabele?

6.2.5



De variatiebreedte (of reikwijdte) wordt niet zo vaak als maat van spreiding gebruikt. Statistici zeggen namelijk dat deze maat “gevoelig is voor uitbijters”. Kan je zelf aangeven wat statistici hiermee precies bedoelen?

6.2.6 Om aan het probleem dat wordt aangekaart in 6.2.5 tegemoet te komen, wordt vaak gebruik gemaakt van de **interkwartielafstand** of de **interdeciaafstand**.

De interkwartielafstand is niets meer dan het verschil tussen het derde en het eerste kwartiel.

De afstand tussen het derde en het eerste kwartiel (of 75^{ste} en 25^{ste} percentiel) geeft doorgaans een realistischer beeld van spreiding dan de variatiebreedte. De variatiebreedte geeft de afstand tussen het minimum en het maximum. Indien we te maken hebben met een dataset waarin een beperkt aantal uitbijters aanwezig is, dan krijgen we via de variatiebreedte een correct beeld van de afstand tussen de extremen, maar dan weten we nog niet waartussen de meeste observaties zich bewegen.

In het geval van de 22 observaties van leeftijd uit 6.1.12 bedraagt de variatiebreedte 28 jaar.

De afstand tussen de 25% jongste respondenten en de 25% oudste respondenten, de interkwartielafstand (K3-K1), bedraagt negen jaar. Deze spreiding geeft een iets realistischer beeld van de leeftijd van de meeste respondenten. Meerbepaald kunnen we uit de interkwartielafstand afleiden hoe hard de centrale 50% van respondenten verschillen van elkaar.

Interdecilen zijn in feite gelijkaardig. Zo kan je de afstand tussen het eerste en het negende deciel berekenen om aan te geven hoever de centrale 80% van je waarnemingen uit elkaar liggen voor de bewuste variabele.

6.2.7 Eerder zagen we hoe we kwantilen en meer specifiek kwartilen konden opvragen in R via de functie `quantile()`. Om de interkwartielafstand te berekenen kunnen we nu als volgt te werk gaan (toegepast voor de variabele schoolgrootte):

```
> quantile(Pirls1$Schoolgrootte, c(.75), na.rm=TRUE) -
  quantile(Pirls1$Schoolgrootte, c(.25), na.rm=TRUE)
75%
172.5
```

Hieruit kunnen we afleiden dat de centrale 50% van de scholen 172,5 leerlingen uit elkaar liggen. Bij het resultaat krijg je eveneens de vermelding "75%". Dit mag je verder negeren, het heeft geen betekenis.



In het bestand "OLP functies.R" hebben we een nieuwe functie weggeschreven: `interkwartiel()`. Aan de hand van deze functie kan je in één commando de interkwartielafstand opvragen. Hieronder hoe het werkt. De eerste lijn dient om wederom een venster op te roepen waarlangs je het bestand "OLP functies.R" opent.

```
> source(file.choose())
> interkwartiel(Pirls1$Schoolgrootte)
75%
172.5
```

6.2.8



Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het 4^{de} leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

a) Wat is het verschil in aantal pc's tussen de 20% laagst scorende en de 20% hoogst scorende scholen?

b) Wat is de interkwartielafstand voor deze variabele?

6.2.9

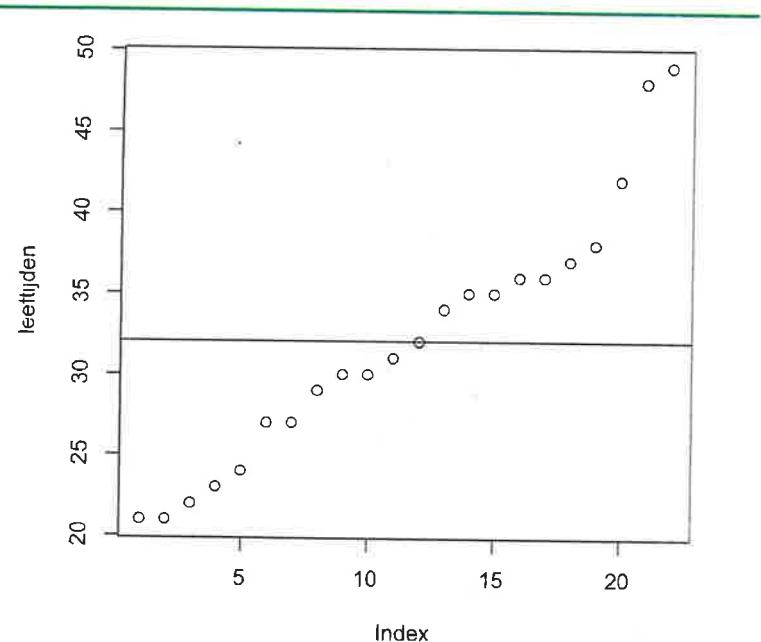


Als we te maken hebben met metingen van tenminste intervalniveau, dan kunnen we afwijkingen tegenover een centraliteitsparameter, zoals het rekenkundig gemiddelde berekenen. Deze afwijkingen kunnen ons een idee geven van de spreiding van observaties.

Hernemen we de 22 leeftijden uit 6.1.12 als illustratie.

21	21	22	23	24	27	27	29	30	30	31	32	34	35	35	36	36	37	38	42	48	49
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

De gemiddelde leeftijd bedraagt 32,14. Onderstaande figuur geeft een weergave van de spreiding van de waarnemingen rond het gemiddelde. De X-as bestaat uit de waarnemingen (van case 1 tot case 22), terwijl de y-as bestaat uit de leeftijden. We hebben een horizontale lijn getrokken op de gemiddelde leeftijd.



Figuur 6.1: Overzicht van de individuele leeftijden uit 6.1.12

Een voor de hand liggende samenvatting van deze afstanden zou het gemiddelde van alle afstanden ten opzichte van het gemiddelde kunnen zijn. Lijkt dat jou een goed idee of is het berekenen van het gemiddelde van die afstanden problematisch? Neem eventueel de proef op de som en reken het na om jezelf te overtuigen.

- 6.2.10**  Louter de afwijkingen tegenover het gemiddelde nemen en die optellen geeft als resultaat nul. Het gemiddelde van deze afstanden is dus geen bruikbare spreidingsmaat.

Een eerste oplossing voor dit probleem is de absolute waarde nemen van de afstand tot het rekenkundig gemiddelde. Vervolgens kunnen we van die absolute afwijkingen een gemiddelde berekenen. Dit levert ons de **gemiddelde absolute afwijking**:

$$\text{Gem.Abs.Afw.} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- 6.2.11** Laat ons deze formule "tot leven" brengen. In R doorlopen we systematisch de verschillende stappen en tonen wat er gebeurt.



STAP 1: de leeftijden ingeven

We kunnen in R gemakkelijk de 22 leeftijden ingeven en dit wegschrijven in een vector met de naam "leeftijden":

```
> Leeftijden<-
c(21,21,22,23,24,27,27,29,30,30,31,32,34,35,35,36,36,37,38,42,
48,49)
```

We maken er een nieuwe dataset van met de naam Lft:

```
> Lft<-data.frame(Leeftijden)
```

STAP 2: afwijkingen ten opzichte van het gemiddelde berekenen

Vervolgens maken we een nieuwe variabele die voor elke respondent de afwijking t.o.v. het gemiddelde bevat. Die nieuwe variabele noemen we Afw_lft en bevat het volgende onderdeel uit de algemene formule van de gemiddelde absolute afwijking:

$$x_i - \bar{x}$$

De onderstaande bewerking in R kunnen we als volgt lezen: trek van elke waarde in de variabele Leeftijden het gemiddelde van de variabele Leeftijden af en schrijf dit weg in de variabele Afw_lft:

```
> Lft$Afw_lft <- Lft$Leeftijden-mean(Lft$Leeftijden)
> Lft
```

	Leeftijden	Afw_lft
1	21	-11.1363636
2	21	-11.1363636
3	22	-10.1363636
4	23	-9.1363636
5	24	-8.1363636
6	27	-5.1363636
7	27	-5.1363636
8	29	-3.1363636
9	30	-2.1363636
10	30	-2.1363636
11	31	-1.1363636
12	32	-0.1363636
13	34	1.8636364
14	35	2.8636364
15	35	2.8636364
16	36	3.8636364
17	36	3.8636364

18	37	4.8636364
19	38	5.8636364
20	42	9.8636364
21	48	15.8636364
22	49	16.8636364

Uit het bovenstaande kunnen we aflezen dat de eerste leeftijd (21 jaar) 11,14 jaar minder is dan de gemiddelde leeftijd. Het is een goede oefening om na te gaan dat de som van de waarden in kolom Afw_lft gelijk is aan nul, net zoals we reeds opmerkten in 6.2.10.

STAP 3: absolute waarde nemen van de afwijkingen

Vervolgens maken we opnieuw een variabele aan die de absolute waarde bevat van de waarden in de variabele Afw_lft. Dit is het volgende deel van de algemene formule van de gemiddelde absolute afwijking:

$$|x_i - \bar{x}|$$

De nieuwe variabele noemen we Abs_afw. Hiervoor maken we gebruik van de functie abs().

```
> Lft$Abs_afw<-abs(Lft$Afw_lft)

> Lft

    Leeftijden      Afw_lft      Abs_afw
 1     21      -11.1363636   11.1363636
 2     21      -11.1363636   11.1363636
 3     22      -10.1363636   10.1363636
 4     23      -9.1363636    9.1363636
 5     24      -8.1363636    8.1363636
 6     27      -5.1363636    5.1363636
 7     27      -5.1363636    5.1363636
 8     29      -3.1363636    3.1363636
 9     30      -2.1363636    2.1363636
10     30      -2.1363636    2.1363636
11     31      -1.1363636    1.1363636
12     32      -0.1363636    0.1363636
13     34      1.8636364    1.8636364
14     35      2.8636364    2.8636364
15     35      2.8636364    2.8636364
16     36      3.8636364    3.8636364
17     36      3.8636364    3.8636364
18     37      4.8636364    4.8636364
19     38      5.8636364    5.8636364
20     42      9.8636364    9.8636364
21     48     15.8636364   15.8636364
22     49     16.8636364   16.8636364
```

STAP 4: gemiddelde berekenen

Een laatste stap bestaat eruit het gemiddelde te berekenen van deze laatst gecreëerde variabele. We tellen met andere woorden alle waarden in Abs_afw op

$$\sum_{i=1}^n |x_i - \bar{x}|$$

en delen dit door n

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Dit kan in R gemakkelijk via de functie mean().

```
> mean(Lft$Abs_afw)
[1] 6.23967
```

Het is een hele klus om via deze tussenstappen de gemiddelde absolute afwijking te berekenen voor een variabele. Daarom hebben we zelf een functie geschreven in het bestand "OLP functies.R" waarmee we in een handomdraai dit kengetal kunnen opvragen. Deze functie heet gemabsafw(). Als we dit toepassen (eerst de file "OLP functies.R" laden via source(file.choose())) krijgen we:

```
> gemabsafw(Lft$Leeftijden)
[1] 6.23967
```

6.2.12

 Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het 4^{de} leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

a) Bereken aan de hand van de verschillende tussenstappen zelf de gemiddelde absolute afwijking voor deze variabele.

(Let op: de eerste stap uit 6.2.11 is hierbij overbodig, we hebben al te maken met een bestaand databestand)

b) Ga na of je goed te werk bent gegaan in a). Hanteer de functie gemabsafw() om je uitkomst bij a) te controleren.

- 6.2.13**  De gemiddelde absolute afwijking heeft als voordeel dat deze eenvoudig kan worden berekend en een zekere overzichtelijkheid geeft. Indien men in een onderzoek niet verder gaat dan een beschrijving van de spreiding van een variabele, dan zou deze maat kunnen volstaan.

Ze is echter niet zo gemakkelijk interpreteerbaar en biedt verder geen voordelen indien we overstappen naar steekproefonderzoek en inferentiële statistiek. Wat we hierna gaan bespreken (de variantie en de standaardafwijking) zijn "de standaard" kengetallen om spreiding weer te geven. Later wordt duidelijker waarom.

Indien we geïnteresseerd zijn in de afwijking ten aanzien van het gemiddelde dan kunnen we ook het probleem van de negatieve verschillen wegwerken door de afstand tot het gemiddelde te kwadrateren. Kwadraattermen zijn immers steeds positief.

De **variantie** berust op dit principe en neemt het kwadraat van de afstand tot het gemiddelde. De variantie wordt vaak aangeduid met σ^2 .

Er zijn twee wijzen voor het berekenen van de variantie. Het is belangrijk om bewust te zijn van het onderscheid tussen beide.

- Een eerste manier om de variantie te berekenen maakt gebruik van het aantal observaties in de noemer. De formule ziet er dan als volgt uit:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Deze formule wordt toegepast als je beschikking hebt over de gegevens van de gehele populatie. Als we bijvoorbeeld gegevens hebben over het inkomen van alle werkende Vlamingen dan kunnen we de variantie in inkomen voor de Vlamingen berekenen aan de hand van bovenstaande formule.

- Het is echter zeer uitzonderlijk dat we beschikken over alle gegevens in de populatie. In de grote meerderheid van de studies doen we berop op steekproeven. Statistici hebben aangetoond dat het bij steekproefgegevens "juister" is om te delen door $n - 1$. Hieronder de formule:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Dus, indien we maar beschikken over gegevens van 1000 werkende Vlamingen (een steekproef) dan delen we de som van de gekwadrateerde afwijkingen door $1000 - 1$ (999) i.p.v. door 1000. In de praktijk maakt dit vaak weinig uit (zeker bij grote steekproeven). Echter, alle statistische pakketten bevatten standaard functies om de variantie te berekenen waarbij gebruik gemaakt wordt van de tweede formule. In wat volgt zullen we dan ook telkens terugrijpen naar de laatste wijze voor het berekenen van varianties.

6.2.14



De berekening van de variantie ziet er nagenoeg hetzelfde uit als de gemiddelde absolute afwijking, maar in plaats van de absolute waarde te nemen van de afwijkingen ten aanzien van het gemiddelde worden nu de kwadraten genomen. In R doorlopen we dan ook enkele gelijkaardige stappen. We bouwen in het onderstaande verder op wat we al in 6.2.11 hebben opgebouwd (de eerste twee stappen) als het gaat om het voorbeeld van de leeftijden.

STAP 1 en 2: zie 6.2.11

STAP 3: de afwijkingen kwadrateren

In 6.2.11 berekenden we voor elke respondent de afwijking van zijn leeftijd t.o.v. de gemiddelde leeftijd. Dit is weggeschreven in de variabele Afw_lft. Nu nemen we daarvan het kwadraat. Om een kwadraat in R te berekenen typen we 2 . De nieuwe variabele noemen we Afw_gekwadr

```
> Lft$Afw_gekwadr<-Lft$Afw_lft^2
> Lft
```

	Leeftijden	Afw_lft	Afw_gekwadr
1	21	-11.1363636	124.01859504
2	21	-11.1363636	124.01859504
3	22	-10.1363636	102.74586777
4	23	-9.1363636	83.47314050
5	24	-8.1363636	66.20041322
6	27	-5.1363636	26.38223140
7	27	-5.1363636	26.38223140
8	29	-3.1363636	9.83677686
9	30	-2.1363636	4.56404959

10	30	-2.1363636	4.56404959
11	31	-1.1363636	1.29132231
12	32	-0.1363636	0.01859504
13	34	1.8636364	3.47314050
14	35	2.8636364	8.20041322
15	35	2.8636364	8.20041322
16	36	3.8636364	14.92768595
17	36	3.8636364	14.92768595
18	37	4.8636364	23.65495868
19	38	5.8636364	34.38223140
20	42	9.8636364	97.29132231
21	48	15.8636364	251.65495868
22	49	16.8636364	284.38223140

STAP 4: de gekwadrateerde afwijkingen optellen en delen door $n - 1$

Bij de gemiddelde absolute afwijking namen we simpelweg het gemiddelde van de nieuw gemaakte kolom (met daarin de absolute waarden van de afwijking). Om de variantie te berekenen kunnen we dit niet toepassen (tenzij we met gegevens voor de hele populatie werken), immers we moeten de som delen door $n - 1$ en niet door n . Hieronder berekenen we in de eerste regel de som a.d.h.v. de `sum()` functie (let op de `na.rm=TRUE`) en schrijven het resultaat weg onder de naam "kwadratensom". Daarna delen we in de tweede regel dit nieuwe object door $n - 1$ door gebruik te maken van de `length()` functie (met daarin ingebed de `na.omit()` functie). Bij die tweede regel is het belangrijk om de haakjes juist te hebben.

```
> Kwadratensom<-sum(Lft$Afw_gekwardr, na.rm=TRUE)
> Kwadratensom/(length(na.omit(Lft$Afw_gekwardr))-1)
[1] 62.59957
```

De variantie in leeftijden bedraagt 62,60. Dit bekwamen we door de formule van s^2 stapsgewijs uit te rekenen. In de praktijk is er in R een functie ingebouwd die rechtstreeks voor ons de variantie berekent: `var()`. Bij deze functie dien je wederom mee te geven aan R dat de NA's verwijderd mogen worden (`na.rm=TRUE`). Deze functie toegepast geeft:

```
> var(Lft$Leeftijden, na.rm=TRUE)
[1] 62.59957
```



6.2.15 Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het vierde leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

a) Bereken aan de hand van de verschillende tussenstappen zelf de variantie voor deze variabele.

(Let op: een aantal stappen heb je mogelijk al gedaan in het teken van 6.2.12)

b) Ga na of je goed te werk bent gegaan in a). Hanteer de in R ingebouwde functie om je uitkomst bij a) te controleren.

6.2.16



Het berekenen van een variantie is één zaak, de interpretatie ervan is iets anders. Wat wil een variantie van 62,6 bij ons voorbeeld van leeftijden eigenlijk zeggen? Hoe moeten we dit interpreteren?

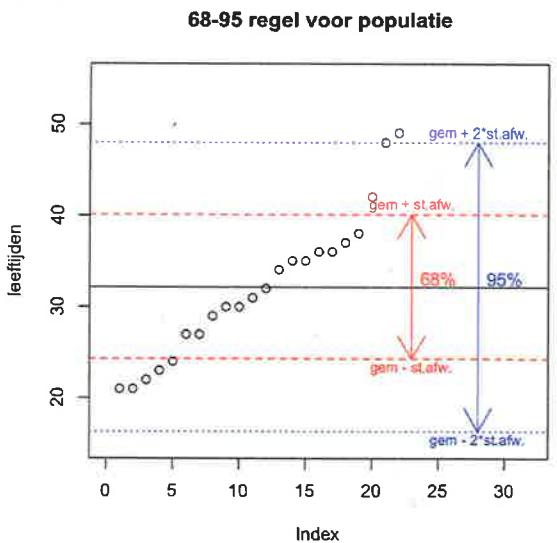
De variantie wordt **niet** uitgedrukt in dezelfde meeteenheid als de observaties. De waarnemingen gaan over leeftijden, terwijl de variantie (leeftijd)² geeft. Dit vereenvoudigt de interpretatie zeker niet.

De **standaardafwijking** is een kengetal dat de variantie herleidt tot een spreidingsmaat in de oorspronkelijke meeteenheid. Het is simpelweg de vierkantswortel van de variantie. De onderstaande formule geeft dit formeel weer, gebruik makend van de formule van de variantie voor steekproefgegevens:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\sigma^2}$$

De standaardafwijking is een kengetal dat zeer vaak wordt gerapporteerd en later in deze bijdrage nog vaak de revue zal passeren. De reden daarom ligt in het feit dat er een aantal eenvoudige regeltjes afgeleid zijn op basis van dit kengetal. Deze regeltjes hebben telkens te maken met "inferentie", dit is conclusies trekken over de populatie aan de hand van steekproefgegevens. Later meer daarover. Nu kunnen we alvast meegeven dat voor een groot aantal variabelen de 68-95-regel opgaat:

- 68% van de eenheden in de populatie ligt binnen het bereik van één standaardafwijking boven en onder het gemiddelde;
- 95% van de eenheden in de populatie ligt binnen het bereik van twee standaardafwijkingen boven en onder het gemiddelde.



Figuur 6.2: Illustratie van de ruwe 68-95 regel

De bovenstaande figuur vat die regel samen. Op basis van onze steekproef kunnen we afleiden dat ongeveer **68% van de leden van de populatie** (waaruit onze steekproef getrokken is) ongeveer tussen de 24 en 40 jaar oud is. 95% van de leden van de populatie is tussen de 16 en 48 jaar oud. Later zullen we zien dat deze conclusie enkel geldig is indien aan een aantal voorwaarden voldaan is.

6.2.17

 De berekening van de standaardafwijking is vrij eenvoudig als je de variatie al kan berekenen (zie 6.2.14). De functie `sqrt()` (komt van squareroot) neemt de wortel uit een getal in R. Bijgevolg is de berekening van de standaardafwijking als volgt te bekomen in R:

```
> sqrt(var(Lft$Leeftijden, na.rm=TRUE))
[1] 7.911989
```

In R is echter ook een ingebouwde functie opgenomen om de standaardafwijking te berekenen: `sd()` (komt van standard deviation). Hieronder de toepassing. Hou bij het gebruik van deze functie wederom rekening met het feit dat je bij ontbrekende waarden het argument `na.rm=TRUE` moet hanteren.

```
> sd(Leeftijden, na.rm=TRUE)
[1] 7.911989
```

6.2.18

 Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het vierde leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

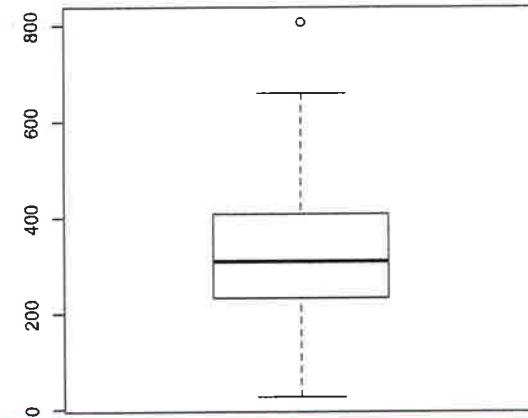
- Bereken de standaardafwijking voor deze variabele.
- Stel dat we gebruik mogen maken van de 68-95 regel, wat kan je dan leren uit de standaardafwijking?

6.3. Grafische weergave van ligging en spreiding: de boxplot

6.3.1

 We hebben eerder in dit boek een aantal grafische voorstellingen van variabelen besproken. Hier gaan we in op een specifieke grafiek die ons in staat stelt om zowel de ligging als de spreiding van een kwantitatieve variabele te visualiseren: de boxplot.

Deze grafiek werd door de statisticus Tukey (1971) geïntroduceerd. Hieronder een voorbeeld van zo'n grafiek voor de variabele schoolgrootte uit ons Pirls1 databestand.



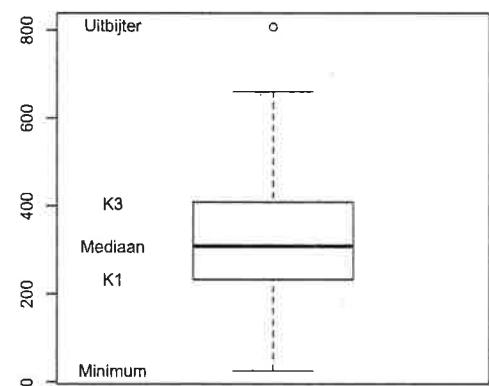
Figuur 6.3: Een voorbeeld van een boxplot (Schoolgrootte uit Pirls1.rda)

Deze grafiek bevat de volgende elementen:

- De dikkere lijn in het midden van de doos geeft de **mediaan**-waarde aan. Zo krijg je ook zicht op de ligging van de variabele.

- Een “doos”, de doos geeft de **interkwartielafstand** aan. De bovenste rand van de doos geeft dus de K3 waarde, de onderste rand de K1-waarde. Dus, hoe hoger de doos, hoe groter de spreiding.
- Vanuit de doos vertrekken twee gestippelde lijnen. Deze worden vaak de snorharen (Whiskers in het Engels) genoemd. Om de lengte van de snorharen te bepalen zijn er twee mogelijkheden. We lichten het toe voor de snorharen naar boven toe, maar een analoge redenering gaat op voor de snorharen naar beneden toe:
 - Eerst wordt berekend welke meetwaarde overeenstemt met het 3^{de} kwartiel plus 1,5 keer de interkwartielafstand. Voor de schoolgrootte weten we dat het derde kwartiel overeenstemt met 405,5 en dat de interkwartielafstand 172,5 bedroeg. Dus de bovengrens wordt berekend als $405,5 + (1,5 \cdot 172,5) = 664,25$.
 - Als alle observaties binnen die bovengrens blijven, dan loopt de bovenste snorhaar tot de maximum geobserveerde waarde.
 - Als niet alle observaties binnen die bovengrens vallen (zoals hier het geval is), dan wordt de bovenste snorhaar getekend tot de verste observatie die niet groter is dan deze bovengrens.
- In het geval er waarden zijn die buiten de boven- en ondergrens vallen dan worden die observaties afzonderlijk met een bolletje weergegeven in de grafiek. De ondergrens stemt overeen met het 1ste kwartiel minus 1,5 keer de interkwartielafstand.

In de figuur hieronder worden de verschillende elementen uit de voorbeeld-boxplot benoemd.



Figuur 1.4: Geannoteerd voorbeeld van een boxplot (Schoolgrootte uit Pirls1.rda)

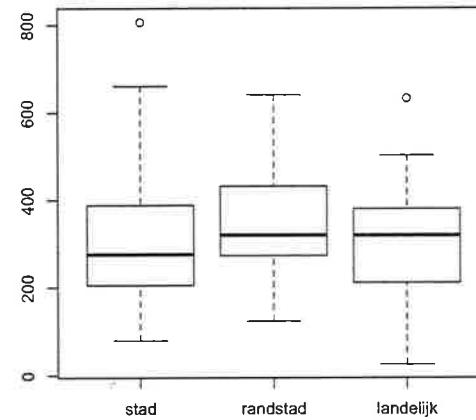
- ### 6.3.2
- Een boxplot oproepen in R is een vrij eenvoudige klus. Simpelweg het commando `boxplot()` gebruiken. Hieronder het commando dat de boxplot produceert uit het voorbeeld.

```
> boxplot(Pirls1$Schoolgrootte)
```

- ### 6.3.3
- Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het 4^{de} leerjaar basisonderwijs. Deze variabele heeft de naam "Pc".

Maak een boxplot aan voor deze variabele en interpreteer deze.

- ### 6.3.4
- De `boxplot()` functie is vrij eenvoudig uit te breiden. Zo zou het wel eens interessant kunnen zijn om zowel de ligging als de spreiding in schoolgrootte van drie verschillende soorten scholen met elkaar te vergelijken: scholen in de stad, scholen in de randsteden en scholen gelegen in ruraal gebied. Dit zou je kunnen doen door drie afzonderlijke boxplots te maken voor telkens een subgroep in het databestand. We kunnen in R deze drie boxplots in een grafiek opvragen. Hieronder het resultaat. In een oogwenk krijgen we zicht op de verschillen tussen de drie types van scholen als het gaat om schoolgrootte.



Figuur 6.5: Boxplot van schoolgrootte alnaargelang de ligging van de school

Om zo'n grafiek aan te maken voegen we ~Pirls1\$Stedelijkheid toe aan het commando. Generieker gesteld, voor het tilde-teken zetten

we de variabele waarvoor we een boxplot willen aanmaken, na het tilde-teken (~) zetten we de categorische variabele die we op de X-as willen plaatsen (op Mac vind je dit teken door Alt+n).

Toegepast om de bovenstaande figuur aan te maken geeft dit:

```
> boxplot(Pirls1$Schoolgrootte~Pirls1$Stedelijkheid)
```

6.3.5



Het databestand Pirls1.RData bevat een variabele die per school aangeeft hoeveel pc's er ter beschikking staan van leerlingen in het vierde leerjaar basisonderwijs. Deze variabele heeft de naam "Pc". Daarnaast is er een variabele die de scholen in 4 categorieën indeelt: scholen met 0 tot 10% anderstaligen, scholen met 11 tot 25% anderstaligen, scholen met 26 tot 50% anderstaligen en scholen met meer dan 50% anderstaligen. Deze variabele heet Pr_anderstalig.

Maak een figuur waarin voor elk van deze types van scholen een afzonderlijke boxplot is gepresenteerd.

Responsen

Respons 6.1.3

Een eenvoudige manier om de modus te achterhalen is het hanteren van een frequentietabel. Dit kan via het volgende commando in R:

```
> table(Pirls1$Stedelijkheid)
```

Hieronder het resultaat:

stad	randstad	landelijk
28	46	57

Hieruit blijkt dat de meeste scholen landelijk gelegen zijn. De modus is m.a.w. de categorie 'landelijk'.

Respons 6.1.5

Herinner je dat je in R een frequentietabel kan opvragen via de functie freqtafel() uit het bestand "OLP Functies.R". We overlopen nog eens de werkwijze:

Om het bestand OLP Functies.R te openen typ je:

```
> source(file.choose( ))
```

Dit opent een venster om een file op je eigen pc te kiezen. Ga naar de map waar je deze file zelf hebt neergezet en kies die file.

Vervolgens roep je een frequentietabel op door het volgende commando:

```
> freqtafel(Pirls1$Samen)
```

Hieronder het resultaat:

X	Freq	Percentage	CumulativeN	CumulativePerc
1 minstens 2-3 keer per week	4	3.030303	4	3.030303
2 een keer in de week	50	37.878788	54	40.909091
3 een keer per maand	52	39.393939	106	80.303030
4 minder dan een keer in de maand	23	17.424242	129	97.727273
5 nooit	3	2.272727	132	100.000000

Aan de hand van de laatste kolom (cumulatieve relatieve frequenties) kunnen we afleiden waar de mediaan zich bevindt. Bijna 41% heeft categorie twee of lager gekozen. Kijken we naar categorie drie dan zien we dat 80,3% deze categorie of een lagere

categorie heeft gekozen. De middelste observatie (cumulatief relatief percentage van 50%) valt dus in deze categorie drie. Bijgevolg is de mediaan categorie drie.

Respons 6.1.9

a) Voor hoeveel scholen kennen we het aantal pc's niet?

Met deze vraag willen we in feite achterhalen wat het aantal "ontbrekende waarden" is voor de variabele "pc". In 6.1.8 leerden we via de functie `length()` nagaan wat respectievelijk het totaal aantal observaties is en het totaal aantal valide observaties (zonder NA's). We passen dit hieronder toe:

```
> length(Pirls1$Pc)
[1] 137
> length(na.omit(Pirls1$Pc))
[1] 129
```

In totaal zijn er dus 137 scholen in het databestand. Van 129 scholen kennen we wel het aantal pc's. Bijgevolg zijn er acht scholen waarvan we het aantal pc's niet kennen. In feite hadden we beide commando's ook in een commando kunnen combineren. Dit kan handiger zijn als je met grotere getallen te maken hebt waarbij hoofdrekenen niet zo evident is. We passen het hieronder toe bij wijze van voorbeeld:

```
> length(Pirls1$Pc) - length(na.omit(Pirls1$Pc))
[1] 8
```

b) Hoeveel pc's zijn er aanwezig in alle geobserveerde scholen samen?

Hiertoe maken we gebruik van de `sum()` functie. Nu we weten dat er ook acht scholen zijn zonder geldige waarnemingen, moeten we rekening houden met deze NA's in onze functie. Hieronder het resultaat:

```
> sum(Pirls1$Pc, na.rm=TRUE)
[1] 1635
```

c) Wat is het gemiddeld aantal pc's in onze steekproef?

We hebben in feite alle gegevens om dit gemiddelde te berekenen: $1635/129 = 12,67$. Willen we het in R doen zoals in 6.1.8 uitgewerkt, dan doen we het als volgt:

```
> sum(Pirls1$Pc, na.rm=TRUE)/length(na.omit(Pirls1$Pc))
[1] 12.67442
```

Gemiddeld zijn er dus in de scholen uit onze steekproef 12 à 13 pc's aanwezig voor leerlingen uit het vierde leerjaar.

Respons 6.1.14

Al deze vragen kunnen we in feite aan de hand van één commando in R beantwoorden:

```
> quantile(Pirls1$Pc, c(.10, .25, .32, .50, .75, .90), na.rm=TRUE)
```

10%	25%	32%	50%	75%	90%
2.00	4.00	5.96	12.00	19.00	25.60

a) D1 = 2

Hiervoor maakten we gebruik van .10 in het `c()` argument.

b) K3 = 19

Hiervoor maakten we gebruik van .75 in het `c()` argument.

c) D9 = 25,6

Hiervoor maakten we gebruik van .90 in het `c()` argument.

d) K2 = 12

Hiervoor maakten we gebruik van .50 in het `c()` argument.

e) Wat is het maximum aantal pc's in de 25% laagst scorende scholen voor deze variabele?

Het antwoord is 4.

Hiervoor maakten we gebruik van .25 in het `c()` argument.

f) Wat is het maximum aantal pc's in de 32% laagst scorende scholen voor deze variabele?

Het antwoord is 5,96 of afgerond 6.

Hiervoor maakten we gebruik van .32 in het `c()` argument.

Respons 6.2.4

a) Om het minimum en het maximum te achterhalen kunnen we het snelst gebruik maken van de `range()` functie in R:

```
> range(Pirls1$Pc, na.rm=TRUE)
[1] 2 60
```

Het minimum aantal pc's beschikbaar is 2, het maximum 60.

b) De variatiebreedte kunnen we hier makkelijk uit het hoofd rekenen: 58.

In R konden we gebruik maken van de functie `reikwijdte()` uit het "OLP Functies.R" bestand.

```
> reikwijdte(Pirls1$Pc)
[1] 58
```

Respons 6.2.5

Met de uitspraak "de reikwijdte is een maat die gevoelig is voor vertekening" bedoelen de statistici dat deze maat mogelijk een slecht beeld geeft van de werkelijke verschillen. Dit is niet zo verwonderlijk. Stel je maar voor dat je te maken hebt met een toevallig zeer vertekende waarneming. Bijvoorbeeld: stel er is één lagere school waar, onder invloed van bepaalde contextkenmerken (ligging op een grote campus met ook een grote secundaire school met vele pc-klassen), 130 pc's ter beschikking staan van leerlingen in het 4^{de} leerjaar. Die ene school zou de reikwijdte uit 6.2.4 optrekken van 58 naar 128. Enkel dit rapporteren wekt de indruk dat er zeer grote verschillen zijn tussen scholen terwijl dit tamelijk overdreven is.

Respons 6.2.8

a) We willen dus de afstand weten dus D8 en D2. Dit kunnen we berekenen via het volgende commando:

```
> quantile(Pirls1$Pc, c(.80), na.rm=TRUE) - quantile(Pirls1$Pc, c(.20),
na.rm=TRUE)
80%
16
```

b) In R konden we gebruik maken van de functie `interkwartiel()` uit het "OLP Functies.R" bestand.

```
> interkwartiel(Pirls1$Pc)
75%
15
```

Respons 6.2.9

Zomaar het gemiddelde berekenen van alle afstanden is niet zo zinvol. We werken het hieronder uit voor de 22 leeftijden. De eerste kolom bevat de leeftijden zelf. In de tweede kolom hebben we voor elke leeftijd het gemiddelde afgetrokken. Willen we het gemiddelde berekenen van deze afwijkingen dan moeten we al die afwijkin-

gen optellen en het resultaat vervolgens delen door het aantal observaties. Het is echter zo dat de som van alle afwijkingen gelijk is aan nul. Dat is niet verwonderlijk. Dat is nu eenmaal de eigenschap van een rekenkundig gemiddelde: zodra we het rekenkundig gemiddelde van leeftijd berekenen dan krijgen we als resultaat die leeftijd waarvoor we weten dat de som van alle afwijkingen ten aanzien van die leeftijd nul is.

Leeftijd	Leeftijd – gemiddelde leeftijd (32,14)
21	-11,14
21	-11,14
22	-10,14
23	-9,14
24	-8,14
27	-5,14
27	5,14
29	-3,14
30	-2,14
30	2,14
31	-1,14
32	-0,14
34	1,86
35	2,86
35	2,86
36	3,86
36	3,86
37	4,86
38	5,86
42	9,86
48	15,86
49	16,86
	som=0

Respons 6.2.12

a) Eerst berekenen we de afwijking van alle observaties t.o.v. het gemiddelde. Met de tweede lijn laten we de eerste 10 obervaties in ons databestand zien voor beide variabelen (de 9^{de} en de 15^{de} kolom in de dataset). Eigenlijk is de eerste lijn de essentie van de eerste stap:

```
> Pirls1$Afw_pc<-Pirls1$Pc-mean(Pirls1$Pc, na.rm=TRUE)
> Pirls1[1:10,c(9,15)]
   Pc      Afw_pc
1  8     -4.674419
2  5     -7.674419
```

```

3     8      -4.674419
4    10      -2.674419
5    20      7.325581
6    2      -10.674419
7   NA       NA
8     3      -9.674419
9     2      -10.674419
10    2      -10.674419

```

Daarna nemen we de absolute waarde en schrijven dit weg in de kolom abs_afw:

```

> Pirls1$Abs_afw<-abs(Pirls1$Afw_pc)
> Pirls1[1:10,c(9,15,16)]

      Pc      Afw_pc      Abs_afw
1     8      -4.674419      4.674419
2     5      -7.674419      7.674419
3     8      -4.674419      4.674419
4    10      -2.674419      2.674419
5    20      7.325581      7.325581
6    2      -10.674419     10.674419
7   NA       NA           NA
8     3      -9.674419      9.674419
9     2      -10.674419     10.674419
10    2      -10.674419     10.674419

```

Tot slot berekenen we het gemiddelde van de laatste kolom:

```

> mean(Pirls1$Abs_afw, na.rm=TRUE)
[1] 7.693889

```

b) We kunnen gebruik maken van de functie `gemabsafw()` uit het "OLP Functions.R" bestand om na te gaan of we goed te werk zijn gegaan.

```

> gemabsafw(Pirls1$Pc)
[1] 7.693889

```

Respons 6.2.15

a) We overlopen de hele berekeningsprocedure. Een deel daarvan overlapt met wat we in 6.2.11 deden.

Eerst berekenen we de afwijking van alle observaties t.o.v. het gemiddelde. Met de tweede lijn laten we de eerste 10 obervaties in ons databestand zien voor beide variabelen (de 9^{de} en de 15^{de} kolom in de dataset). Eigenlijk is de eerste lijn de essentie van de eerste stap:

```

> Pirls1$Afw_pc<-Pirls1$pc-mean(Pirls1$Pc, na.rm=TRUE)
> Pirls1[1:10,c(9,15)]

      Pc      Afw_pc
1     8      -4.674419
2     5      -7.674419
3     8      -4.674419
4    10      -2.674419
5    20      7.325581
6     2      -10.674419
7   NA       NA
8     3      -9.674419
9     2      -10.674419
10    2      -10.674419

```

Daarna nemen we het kwadraat van die afwijkingen en schrijven dit weg in de kolom afw_gekwadr:

```

> Pirls1$Afw_gekwadr<-Pirls1$Afw_pc^2
> Pirls1[1:10,c(9,15,16)]

      Pc      Afw_pc      Afw_gekwadr
1     8      -4.674419      21.850189
2     5      -7.674419      58.896701
3     8      -4.674419      21.850189
4    10      -2.674419      7.152515
5    20      7.325581      53.664143
6     2      -10.674419     113.943213
7   NA       NA           NA
8     3      -9.674419      93.594375
9     2      -10.674419     113.943213
10    2      -10.674419     113.943213

```

Let op: mogelijk krijg je niet wat je wenst bij de tweede regel in het bovenstaande. Als je bijvoorbeeld doorbouwde op de uitkomst in 6.2.11 dan is afw_gekwadr niet de 16de variabele in het databestand, maar de 17de. Bijgevolg moet je de tweede regel aanpassen naar:

```
> Pirls1[1:10,c(9,15,17)]
```

Vervolgens berekenen we de som van al die gekwadrateerde afwijkingen en schrijven dit weg in een vector met de naam "kwadratensom":

```
> Kwadratensom<-sum(Pirls1$Afw_gekwadr, na.rm=TRUE)
```

Tot slot delen we de kwadratensom door het aantal geldige observaties min één om de variantie uit te komen:

```
> Kwadratensom/(length(na.omit(Pirls1$Pc))-1)
[1] 96.89317
```

b) We kunnen gebruik maken van de functie `var()` om na te gaan of we goed te werk zijn gegaan. Let daarbij op het extra argument `na.rm=TRUE` waarmee we aangeven dat NA's verwijderd mogen worden.

```
> var(Pirls1$Pc, na.rm=TRUE)
[1] 96.89317
```

Respons 6.2.18

a) De berekening van de standaardafwijking is eenvoudig in R:

```
> sd(Pirls1$Pc, na.rm=TRUE)
[1] 9.843433
```

b) De berekeningen voor de 68%-regel kan je als volgt doen in R:

```
> mean(Pirls1$Pc, na.rm=TRUE)-sd(Pirls1$Pc, na.rm=TRUE)
[1] 2.830986
> mean(Pirls1$Pc, na.rm=TRUE)+sd(Pirls1$Pc, na.rm=TRUE)
[1] 22.51785
```

De eerste regel berekent de "ondergrens", de derde regel de "bovengrens". 68% van de leden van de populatie (= 68% van de Vlaamse scholen) heeft tussen de 2,8 en 22,5 pc's ter beschikking voor de leerlingen in het 4^{de} leerjaar basisonderwijs.

De berekeningen voor de 95%-regel kan je als volgt doen in R:

```
> mean(Pirls1$Pc, na.rm=TRUE)-2*sd(Pirls1$Pc, na.rm=TRUE)
[1] -7.012447
> mean(Pirls1$Pc, na.rm=TRUE)+2*sd(Pirls1$Pc, na.rm=TRUE)
[1] 32.36128
```

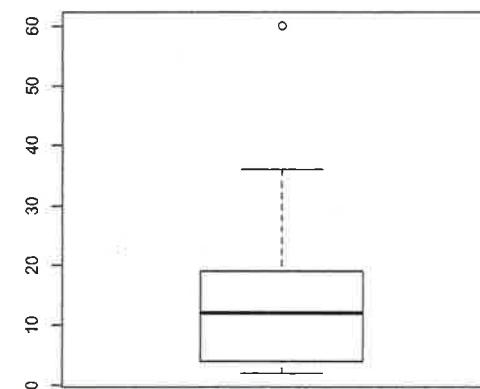
Uit deze toepassing blijkt al vlug het gevaar van het "blindelings" afleiden van gegevens naar de populatie toe op basis van een steekproef. Immers, op basis van de bovenstaande berekening kom je bij een negatief aantal pc's uit. Dit is uiteraard absurd. Later zullen we zien dat het bij deze variabele in feite niet aangewezen is deze regel toe te passen.

Respons 6.3.3

Het commando is vrij eenvoudig:

```
> boxplot(Pirls1$Pc)
```

Dit geeft:



Figuur 6.3.3: Boxplot van het aantal pc's ter beschikking van leerlingen in het 4^{de} leerjaar basisonderwijs

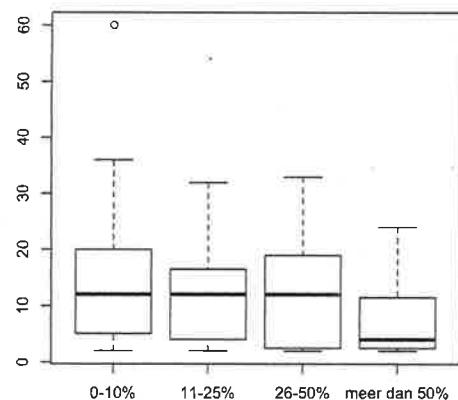
De bovenstaande figuur toont dat er één school is die vrij extreem scoort: 60-tal pc's ter beschikking van de leerlingen in het 4^{de} leerjaar. Daarnaast kunnen we afleiden dat er wel wat verschillen zijn tussen scholen als het gaat om het aantal pc's ter beschikking van de leerlingen. 25% van de scholen heeft meer dan 20 pc's ter beschikking. Maar daarnaast heeft ook 25% van de scholen minder dan 4 pc's ter beschikking.

Respons 6.3.5

Deze figuur maak je aan door middel van het volgende commando:

```
> boxplot(Pirls1$pc~Pirls1$Pr_anderstalig)
```

Dit geeft:



Figuur 6.3.5: Boxplot van het aantal pc's ter beschikking van leerlingen in het 4^{de} leerjaar basisonderwijs naargelang het percentage anderstaligen in de school

Uit deze figuur kunnen we afleiden dat de mediaan voor het aantal pc's in de school duidelijk lager is voor scholen met een hoog percentage allochtonen (meer dan 50%). Bovendien valt op dat deze laatste groep van scholen een kleinere spreiding vertonen. Met andere woorden, deze scholen gelijken meer op elkaar aangaande het aantal pc's dan de scholen met minder anderstalige kinderen.

Gehanteerde functies

Functie	Doelstelling	Bron
<code>abs()</code>	Deze functie neemt de absolute waarde van wat er tussen de haakjes staat. Dit kan een getal zelf zijn, maar ook een variabele. In dat geval wordt de absolute waarde van elke geobserveerde meetwaarde in die variabele.	R basispakket
<code>as.numeric()</code>	Verandert een categorische variabele in een numerieke variabele.	R basispakket
<code>boxplot()</code>	Stelt je in staat om een boxplot te produceren.	R basispakket
<code>gemabsafw()</code>	Berekent de gemiddelde absolute afwijking voor een variabele.	OLP functies.R
<code>interkwartiel()</code>	Berekent de interkwartielafstand voor een variabele.	OLP functies.R
<code>length()</code>	Geeft als resultaat het aantal observaties voor een variabele of een hele dataset.	R basispakket
<code>mean()</code>	Berekent het rekenkundig gemiddelde van een nummerieke variabele.	R basispakket
<code>median()</code>	Berekent de mediaan waarde voor een nummerieke variabele.	R basispakket
<code>max()</code>	Geeft als resultaat de maximum meetwaarde die we observeren voor een variabele.	R basispakket
<code>min()</code>	Geeft als resultaat de minimum meetwaarde die we observeren voor een variabele.	R basispakket
<code>na.omit()</code>	Deze functie verwijdt alle observaties waarvoor we geen geldige waarde hebben voor een welbepaalde variabele, of zelfs een hele set van variabelen.	R basispakket
<code>quantile()</code>	Geeft de gewenste kwantilen als resultaat. Standaard berekent deze functie de 3 kwartielwaarden.	R basispakket
<code>range()</code>	Geeft in één beweging de minimum en de maximum geobserveerde meetwaarden weer.	R basispakket
<code>reikwijdte()</code>	Berekent de reikwijdte voor een variabele.	OLP functies.R
<code>sd()</code>	Berekent de standaardafwijking voor een variabele.	R basispakket
<code>sqrt()</code>	Berekent de vierkantswortel.	R basispakket
<code>sum()</code>	Berekent de som van een reeks getallen. Indien we een variabele plaatsen tussen de haakjes, dan telt deze functie alle meetwaarden in de variabele op met elkaar.	R basispakket
<code>summary()</code>	Een functie die in één beweging enkele samenvattende kengetallen voor een variabele weergeeft: gemiddelde, mediaan en het eerste en derde kwartiel.	R basispakket
<code>var()</code>	Berekent de variantie voor een variabele.	R basispakket

Parameters van vorm

DOELSTELLINGEN:

Na dit hoofdstuk:

- kan je de verschillende kengetallen voor vorm (skewness en kurtosis) uitleggen;
- kan je deze kengetallen via het pakket 'moments' in R berekenen;
- ben je in staat beide kengetallen te interpreteren.

NODIGE FILES:

Opleidingen1.RData

een file met daarin een aantal variabelen over 268 opleidingen die bedienden uit een welbepaalde financiële instelling volgden in de voorbije 5 jaar

NODIGE PAKKETTEN IN R:

moments

*een pakket ontwikkeld om de 'momenten' van een verdeling te berekenen. De kengetallen die we bespreken in dit hoofdstuk zijn toepassingen van de 'momenten' van een verdeling. Dit pakket is geen standaard onderdeel van R. In hoofdstuk 2 leerde je hoe je nieuwe pakketten aan de standaardinstallatie van R moet toevoegen. Meer informatie over het pakket zelf is te vinden op:
<http://rss.acs.unt.edu/Rdoc/library/moments/html/00Index.html>*



In de vorige hoofdstukken kwamen reeds verschillende manieren aan bod om de gegevens die we hebben op een of andere manier te beschrijven of samen te vatten. We bespraken de eerste grove samenvatting aan de hand van frequentietabellen. Vervolgens introduceerden we kengetallen die iets zeggen over de ligging en de spreiding van de waarnemingen voor onze variabelen. In dit hoofdstuk introduceren we de laatste kengetallen om een andere eigenschap van de verdeling van variabelen te beschrijven: de vorm van de verdeling.

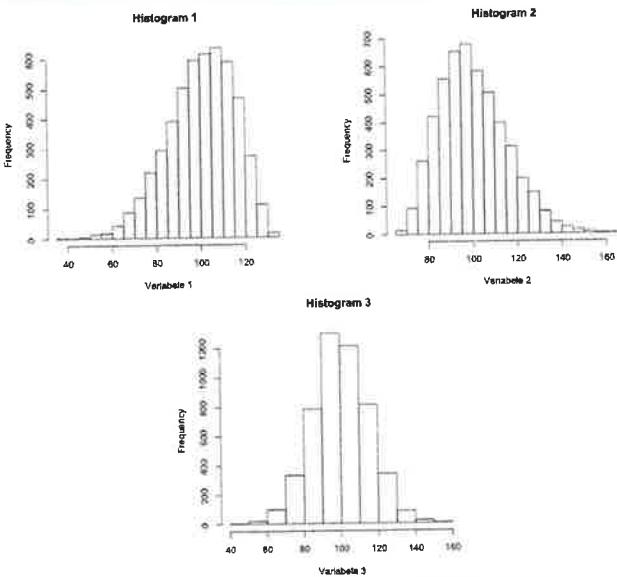
7.1. Scheefheid

7.1.1



In de onderstaande figuur staan 3 histogrammen. Bestudeer deze histogrammen.

- Wat is het meest opvallende verschil bij deze histogrammen?
- Hoe verwacht je voor elk van deze histogrammen dat het rekenkundig gemiddelde en de mediaan zich gaan verhouden? Geef voor elk histogram aan of je verwacht dat het rekenkundig gemiddelde bij benadering gelijk is aan de mediaan, groter is dan de mediaan of lager is dan de mediaan.

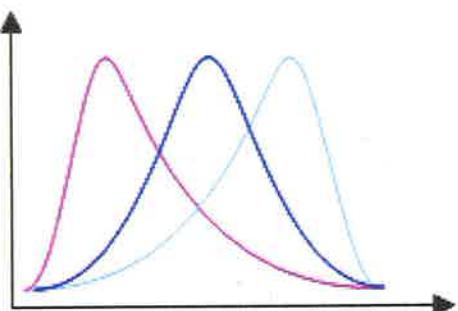


Figuur 7.1: Drie voorbeeldhistogrammen

- 7.1.2** Het grote verschil tussen de drie bovenstaande histogrammen is dat het derde histogram een variabele beschrijft die (bij benadering) **symmetrisch verdeeld** is. De andere twee histogrammen beschrijven een variabele die **scheef verdeeld** is.

We noemen een variabele **rechtsscheef verdeeld** indien de staart langs de rechterzijde langer is. In de onderstaande figuur is dit de roze verdeling. In het Engels noemen we dit *positively skewed*.

Bij een **linksscheef verdeelde** variabele is de staart langs de linkerzijde langer. In de onderstaande figuur gaat het om de lichtblauwe curve. In het Engels is dit *negatively skewed*.



Figuur 7.2: Een rechtsscheve, een symmetrische en een linksscheve verdeling

Verder in dit OLP zullen we stilstaan bij het belang van de scheefheid van de verdeling van een variabele, als we het hebben over de normale verdeling.

- 7.1.3** Op basis van 2 parameters van ligging (mediaan en gemiddelde) kunnen we een belangrijke eigenschap van symmetrische en scheef verdeelde variabelen definiëren. Hoe denk je zelf dat het is? Vul het onderstaande schema aan met '>', '=' of '<'.

rechtsscheef	symmetrisch	linksscheef
mediaan ... gemiddelde	mediaan ... gemiddelde	mediaan ... gemiddelde

7.1.4

- We zouden de regels (eigenschappen) die we in 7.1.3 beschreven, kunnen hanteren om te bepalen of een variabele al dan niet symmetrisch verdeeld is. Wat deze methode niet toelaat is om op een eenduidige wijze te bepalen hoe scheef een variabele verdeeld is.

We kunnen echter ook een kengetal gebruiken dat de scheefheid van de verdeling van een variabele beschrijft. Dit getal wordt in de statistische literatuur de "skewness" van een variabele genoemd. De volgende formule geeft weer hoe deze skewness berekend wordt. We geven ze enkel ter volledigheid, in de praktijk zal je deze formule zelf nooit ter hand nemen.

$$\text{skewness} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{N}/\sigma^3$$

De teller kennen we voor een groot stuk. Deze bevat de som van de afwijkingen van het gemiddelde, nadat die afwijkingen tot de 3^{de} macht verheven zijn. Die som wordt vervolgens gedeeld door het aantal observaties (N). De noemer bevat de standaardafwijking (σ) verheven tot de 3^{de} macht. Voor een perfect symmetrisch verdeelde variabele bedraagt de skewness de waarde nul. Is de waarde groter dan nul, dan wil dit zeggen dat de variabele 'positively skewed' is, of rechtsscheef verdeeld. Een waarde lager dan nul betekent dat de variabele 'negatively skewed' is, of linksscheef verdeeld. Het onderstaande schema vat dit nog eens samen:

rechtsscheef	symmetrisch	linksscheef
skewness > 0	skewness = 0	skewness < 0
mediaan < gemiddelde	mediaan = gemiddelde	mediaan > gemiddelde

7.1.5

- Standaard bevat R geen functie die de scheefheid van een variabele berekent. We kunnen beroep doen op een pakket dat we afzonderlijk moeten installeren: moments. Vergewis je ervan dat dit pakket geïnstalleerd is op je computer. Als het pakket op je computer staat, moet je het in elke nieuwe sessie in R opnieuw opladen. Dit doe je via het library() commando:

> library(moments)

Het moments pakket bevat de functie skewness() die voor een variabele het kengetal voor scheefheid gaat berekenen. Belangrijk daarbij is dat

we bij dat commando opnieuw moeten expliciteren wat er met ontbrekende waarden (NA's) moet gebeuren. Dit doen we via het extra argument na.rm=TRUE. Hieronder dit commando toegepast voor een fictieve variabele (Variabele1) in een fictieve dataset (Dataset1):

```
> skewness(Dataset1$Variabele1, na.rm=TRUE)
```

7.1.6



Voor dit hoofdstuk maken we gebruik van het databestand Opleidingen1.RData. Dit bestand bevat een selectie van gegevens die resulteerden uit een bevraging in een financiële instelling. Men trok een steekproef uit alle bankbedienden die gedurende een periode van 5 jaar minstens één maal een opleiding volgden, al dan niet intern georganiseerd. Aan deze bankbedienden werd vervolgens een vragenlijst opgestuurd met een aantal vragen over die opleiding.

We beschikken hier over een selectie van variabelen voor elke bankbediende in de steekproef. We lichten ze even toe:

- Kostprijs: inschatting van de kostprijs van de opleiding;
- Duur: het aantal uren dat de opleiding duurde;
- Ndeelnemers: het aantal deelnemers dat de opleiding volgde samen met de bediende;
- Ancien: het aantal maanden anciënniteit dat de bankbediende had op het moment van de opleiding;
- Inkomen: het netto-inkomen dat de bankbediende had op het moment van de opleiding;
- Tevredenheid: een algemene tevredenheidscore over de opleiding;
- Nut: een inschatting van de mate waarin de bankbediende de opleiding nuttig vond om z'n huidige takenpakket uit te voeren (gebaseerd op een likertschaal met 6 items, herrekend naar een getal met gemiddelde 0 en standaardafwijking 1).

Ga voor de volgende van deze variabelen in R na in welke mate ze al dan niet scheef verdeeld zijn, interpreteer wat dit inhoudelijk betekent en verifieer eventueel je conclusie a.d.h.v. een histogram: Kostprijs, Duur, Ndeelnemers en Ancien.

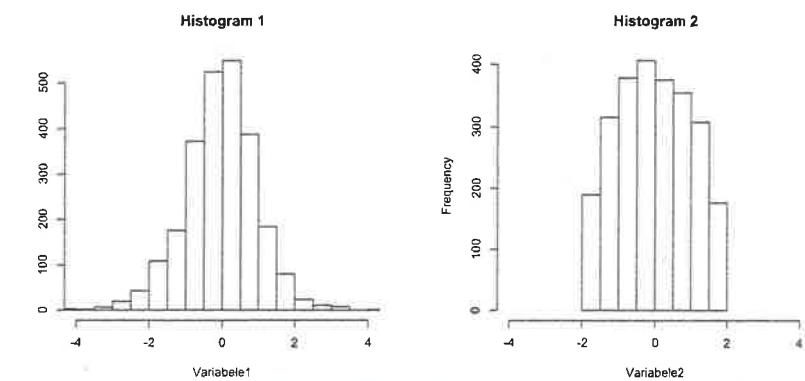
7.2. Platheid (Kurtosis)

7.2.1



In de onderstaande illustratie staan 2 histogrammen. Bestudeer deze histogrammen.

Wat is het meest opvallende verschil tussen beide histogrammen?

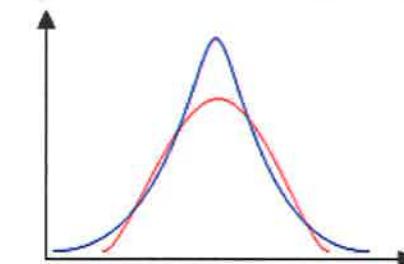


Figuur 7.3: Twee voorbeeldhistogrammen

7.2.2



Het verschil dat we in bovenstaande histogrammen zien is een verschil in 'gepiektheid' of kurtosis in vaktermen. De onderstaande figuur geeft dit verschil op een andere wijze weer.



Figuur 7.4: Voorstelling van het verschil tussen leptokurtic en platykurtic verdelingen

Indien de verdeling van een variabele een meer uitgesproken piek vertoont en 'dikkere uiteinden' heeft dan is de variabele '**leptokurtic**' verdeeld. Is de variabele 'platter verdeeld' met 'minder dikke uiteinden' dan is de variabele '**platykurtic**' verdeeld.

In de bovenstaande figuur is de rode verdeling platykurtic en de blauwe verdeling leptokurtic.

- 7.2.3** Als je beide verdelingen in de bovenstaande figuur bekijkt, hoe verwacht je dan dat de standaardafwijkingen voor beide verdelingen zich verhouden. Is de ene groter dan de andere of niet?

- 7.2.4** Het is geen optie om de kurtosis van de verdeling van een variabele enkel op het oog na te gaan. Daartoe zijn wederom verschillende kengetallen ontwikkeld. We zullen in deze cursus enkel het kengetal gebruiken dat door R wordt gegeven. De formule van dit kengetal lijkt zeer sterk op de formule voor de scheefheid:

$$\text{kurtosis} = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{N} / \sigma^4$$

Het enige verschil met de formule voor scheefheid is dat in de teller de afwijkingen ten opzichte van het gemiddelde tot de 4^{de} macht verheven worden en dat in de noemer de standaardafwijking tot de 4^{de} macht verheven wordt. Dit kengetal heeft niet zoals scheefheid de waarde nul als referentiepunt. Bij kurtosis ligt het referentiepunt bij waarde 3. Een waarde 3 betekent eigenlijk dat de variabele **mesokurtic** verdeeld is. Later komen we terug op wat dat precies wil zeggen, maar onthoud nu reeds dat het referentiepunt in feite de kurtosis is van de normaalverdeling (maar daarover later meer). Een waarde hoger dan 3 wijst op een leptokurtic verdeelde variabele. Is de kurtosis kleiner dan 3 dan wijst dit op een platykurtic verdeelde variabele. Wees er ook attent op dat in sommige statistische softwarepakketten een andere formule wordt gehanteerd bij het berekenen van de kurtosis, namelijk ze trekken van de kurtosis de waarde 3 af zodanig dat het referentiepunt op de waarde nul komt te liggen. In R (zie hieronder) wordt dit niet gedaan in het pakket dat wij voor dit boek han-

teren. Dus in de output uit R zal de waarde 3 wijzen op de kurtosis van een mesokurtic verdeelde variabele. Schematisch samengevat geeft dit:

platykurtic	mesokurtic	leptokurtic
kurtosis < 3	kurtosis = 3	kurtosis > 3

7.2.5

 Om in R een schatting te verkrijgen van de kurtosis van de verdeling van een variabele ga je analoog te werk zoals voor het bepalen van de scheefheid (zie 7.1.5): we maken gebruik van het pakket 'moments'. In dat pakket zit de functie `kurtosis()`. Hieronder passen we dit toe voor een fictieve variabele (`Variabele1`) uit het databestand (`Dataset 1`). Let op het feit dat we opnieuw eerst `moments` moeten laden indien we R tussendoor hebben afgesloten. Let tevens op het extra argument `na.rn=TRUE` dat ervoor zorgt dat de ontbrekende waarden (NA's) eerst verwijderd worden.

```
> library(moments)
> kurtosis(Dataset1$Variabele1, na.rn=TRUE)
```

7.2.6

 Open het databestand `Opleidingen1.RData`. Ga voor de volgende van deze variabelen in R na in welke mate ze al dan niet lepto- of platykurtic verdeeld zijn en interpreteer wat dit inhoudelijk betekent: Kostprijs, Duur, Ndeelnemers en Nut.

Responsen

Respons 7.1.1

a) Histogrammen 1 en 2 zijn minder symmetrisch verdeeld dan histogram 3. Histogram 1 vertoont een langere staart aan de linkerzijde, histogram 2 een langere staart aan de rechterzijde.

b) Bij histogram 1 verwachten we dat het rekenkundig gemiddelde lager is dan de mediaan. De uitschieters links in de verdeling trekken het rekenkundig gemiddelde naar beneden. Van de mediaan weten we dat die niet gevoelig is voor extreemere waarnemingen. Deze zal dus niet worden beïnvloed door de extreemere waarden in de linkse staart.

Om diezelfde reden verwachten we dat bij histogram 2 het rekenkundig gemiddelde hoger zal zijn dan de mediaan en dat bij histogram 3 beide waarden grosso modo gelijk zullen zijn.

Respons 7.1.3

De tabel ziet er als volgt aangevuld uit:

rechtsscheef	symmetrisch	linksscheef
mediaan < gemiddelde	mediaan = gemiddelde	mediaan > gemiddelde

Dit is een formalisering van de redenering die we opbouwden in de respons op opdracht b) uit 7.1.1 (zie bovenstaande respons).

Bij een rechtsscheef verdeelde variabele is de mediaan kleiner dan het gemiddelde. Bij een linksscheve verdeling is de mediaan groter dan het gemiddelde. Een symmetrische verdeling heeft de eigenschap dat de mediaan gelijk is aan het gemiddelde.

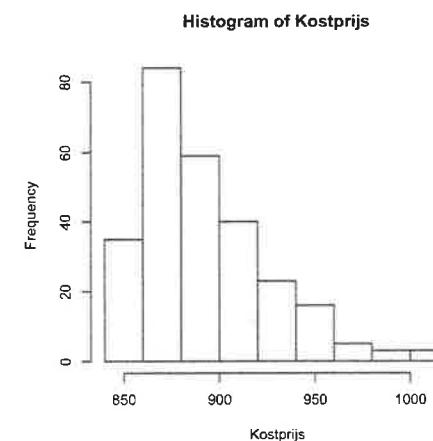
Respons 7.1.6

Hieronder staan de verschillende commando's en de bijhorende output in R om de scheefheid van die bewuste variabelen na te gaan:

```
> skewness(Kostprijs)
[1] 1.155289
> skewness(Duur)
[1] -0.5688765
```

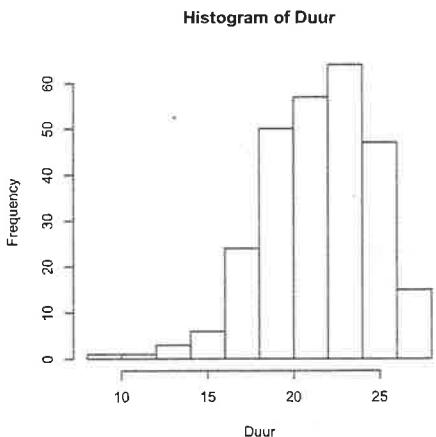
```
> skewness(Ndeelnemers)
[1] 0.1013521
> skewness(Ancien)
[1] -0.8055299
```

Voor de kostprijs van de opleidingen stellen we vast dat de verdeling "positively skewed" is. M.a.w. de verdeling is rechtsscheef verdeeld. De staart loopt uit naar rechts toe. Dit betekent dat er een beperkt aantal opleidingen is die opvallend duurder zijn dan de andere opleidingen. Het volgende histogram illustreert deze tendens:



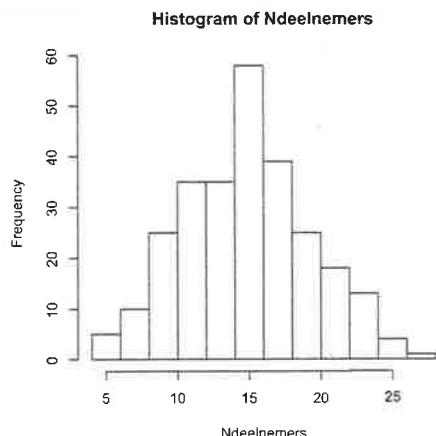
Figuur 7.5: Histogram voor de variabele Kostprijs

Voor de variabele Duur is de skewness negatief (skewness = -0,57). Deze variabele is dus linksscheef verdeeld. De staart is aan de linkse kant langer uitgerekt. Dit betekent dat er een aantal opleidingen zijn die beduidend minder uren in beslag nemen dan het grootste deel van de opleidingen. Dat zien we mooi in het onderstaande histogram.



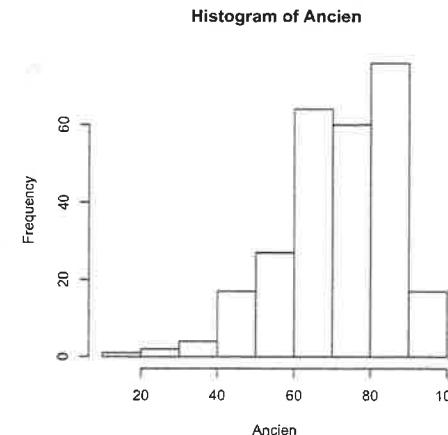
Figuur 7.6: Histogram voor de variabele Duur

De scheefheid voor de variabele Ndeelnemers is ook positief (skewness = 0,10), zij het minder uitgesproken dan bijvoorbeeld bij de variabele kostprijs. We verwachten dus dat de staart van de verdeling aan de rechtse kant langer zal zijn, maar dat deze tendens minder uitgesproken zal zijn dan bij de variabele Kostprijs. Er zijn dus wel enkele opleidingen waarvoor het aantal deelnemers relatief hoog is. Het onderstaande histogram geeft dit ook aan. Maar wat onmiddellijk opvalt aan dit histogram is dat deze tendens inderdaad niet zo uitgesproken is: het lijkt erg op een symmetrische verdeling.



Figuur 7.7: Histogram voor de variabele Ndeelnemers

Tot slot wijst de scheefheid van de variabele Ancien op een "negatively skewed" verdeling (skewness = -0,81). De staart is meer uitgestrekt aan de linkse zijde. Er is een beduidend aantal bankbedienden die weinig maanden anciënniteit hadden op het moment dat ze de opleiding gingen volgen.



Figuur 7.8: Histogram voor de variabele Ancien

Respons 7.2.1

Histogram 1 vertoont een duidelijkere piek in het midden van de verdeling en aan de uiteinden van de verdeling stellen we ook meer waarnemingen vast. Histogram 1 is platter verdeeld, en heeft geen uitlopers aan beide uiteinden zoals histogram 2.

Respons 7.2.3

De standaardafwijking is onafhankelijk van de kurtosis van een verdeling. Je kan dit als volgt zien: bij de variabele die meer leptokurtic verdeeld is (hogere piek en dikdere uiteinden) wordt de standaardafwijking grotendeels bepaald door het relatief groter aantal waarnemingen in de uiteinden. Bij een variabele die meer platykurtic is verdeeld (lagere piek en dunne uiteinden) wordt de standaardafwijking grotendeels bepaald door het grotere aantal waarnemingen rondom het centrum van de verdeling.

Respons 7.2.6

Van twee variabelen is de kurtosis hoger dan de waarde 3: Kostprijs en Duur. Voor die variabelen kunnen we concluderen dat de verdeling eerder leptokurtic is van vorm. Dit houdt in dat er langere uitlopende staarten zijn en een meer uitgesproken piek. Er zijn veel opleidingen die ongeveer even duur zijn, maar ook een aanzienlijk deel opleidingen die erg duur zijn of erg goedkoop zijn. Voor de variabele Kostprijs is deze tendens meer uitgesproken dan voor de variabele Duur.

De twee overige variabelen zijn platykurtic verdeeld: de kurtosis is lager dan 3. Beide variabelen vertonen een minder uitgesproken piek en hebben minder uitlopende staarten aan beide uiteinden. Dit betekent dat er nauwelijks opleidingen als extreem onnuttig of extreem nuttig worden beoordeeld (geen uitlopende staarten), en dat er een groot aantal opleidingen is die minder extreem van elkaar verschillen qua nuttigheid zonder dat er sprake is van een bepaalde nuttigheidsscore die zeer veel opleidingen krijgen.

```
> kurtosis(Opleidingen1$Kostprijs, na.rm=TRUE)
[1] 4.205113
> kurtosis(Opleidingen1$Duur, na.rm=TRUE)
[1] 3.547953
> kurtosis(Opleidingen1$Ndeelnemers, na.rm=TRUE)
[1] 2.687035
> kurtosis(Opleidingen1$Nut, na.rm=TRUE)
[1] 2.621777
```

Gehanteerde functies

Functie	Doelstelling	Bron
kurtosis()	Berekent de kurtosis van de verdeling van een variabele. De waarde 3 wordt als referentiepunt gehanteerd. Een waarde 3 wijst op een mesokurtic verdeelde variabele.	moments
skewness()	Deze functie geeft als resultaat het kengetal voor scheefheid. Een waarde nul wijst op een symmetrisch verdeelde variabele.	moments

HOOFDSTUK 8

De (standaard-)normaalverdeling

DOELSTELLINGEN:

Na dit hoofdstuk:

- weet je wat een kansverdeling is;
- weet je wat de normaalverdeling is;
- weet je wat de standaardnormaalverdeling is;
- weet je wat een z-score is;
- kan je een z-score berekenen in R;
- kan je in R gebruik maken van de eigenschappen van de standaard-normaalverdeling.

NODIGE FILES:

Opleidingen1.RData

OLP Functies.R

een file met daarin aangepaste functies die bij dit OLP horen

8.1. De normaalverdeling

8.1.1

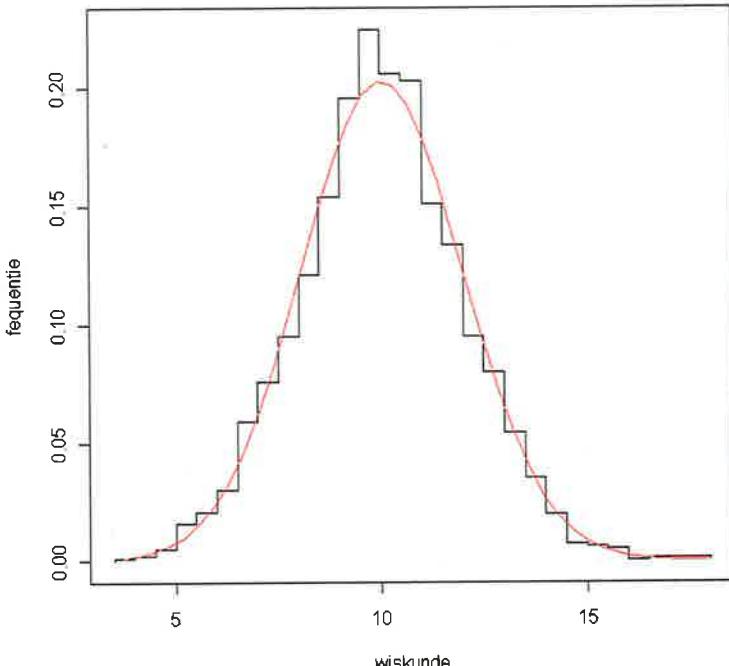


In het vorige hoofdstuk bespraken we de relevante parameters om de vorm van de verdeling van een variabele te beschrijven: de scheefheid en de kurtosis (platheid). Naast het feit dat zowel de scheefheid als de kurtosis op zich informatief zijn, dienen ze in de praktijk van de onderzoeker vaak een andere functie.

Zowel de scheefheid als de kurtosis leren ons in hoeverre de variabele al dan niet normaal verdeeld is. De normaalverdeling is een theoretisch model, een voorbeeld van een verdeling welke we als volgt kunnen definiëren:

Een normale verdeling is een **symmetrische, unimodale, klokvormige** verdeling.

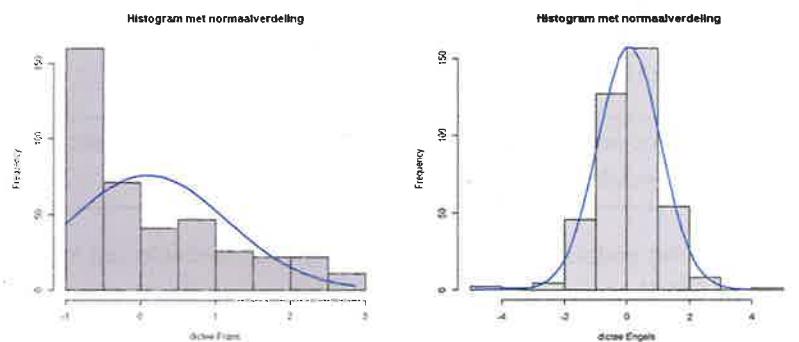
Het onderstaande histogram geeft een voorbeeld van een variabele die (bij benadering) normaal verdeeld is.



Figuur 8.1: Histogram van wiskundescores met normaalverdelingcurve

Bovenop dit histogram hebben we de normaalverdeling in het rood getekend. Indien we dit histogram bekijken dan zien we in feite dat deze niet 100% perfect overeenstemt met de normaalverdeling, maar dat deze grotendeels de vorm van deze curve volgt. Vandaar dat we stellen dat deze variabele 'bij benadering' normaal verdeeld is.

De onderstaande twee histogrammen zijn voorbeelden van variabelen waarvan de verdeling duidelijker afwijkt van de normaalverdeling:



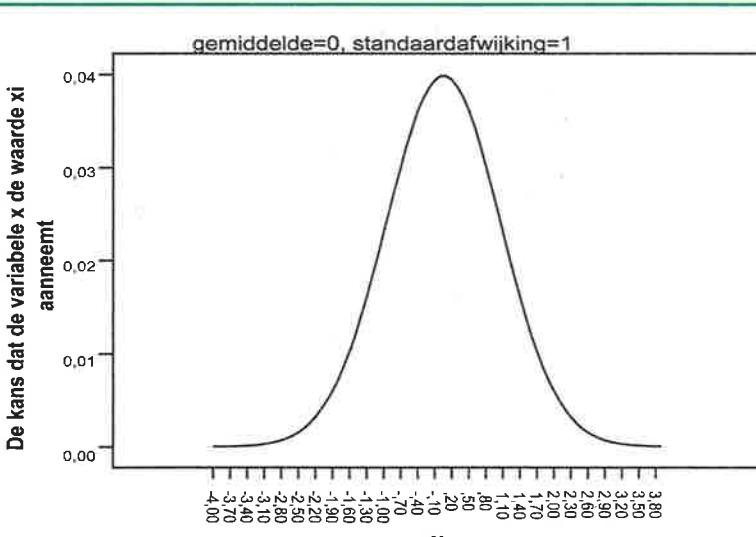
Figuur 8.2: Niet normaalverdeelde variabelen

Waarom is de normaalverdeling zo interessant?

Het antwoord ligt in het feit dat de normaalverdeling een **theoretische kansverdeling** is. Om daar het belang van in te zien staan we eerst stil bij wat een kansverdeling is en wat daar de eigenschappen van zijn.

Een **kansverdeling** is de verdeling van de kans dat een bepaalde waarde van een variabele voorkomt.

Deze verdeling kunnen we visualiseren door alle mogelijke waarden van deze variabele af te zetten tegen de kans dat we deze waarde waarnemen. De onderstaande figuur visualiseert de kansverdeling van een normaal verdeelde variabele met een gemiddelde score 0 en een standaardafwijking 1. Op de x-as staan de mogelijke waarden, de y-as geeft de kans dat deze waarde voorkomt.

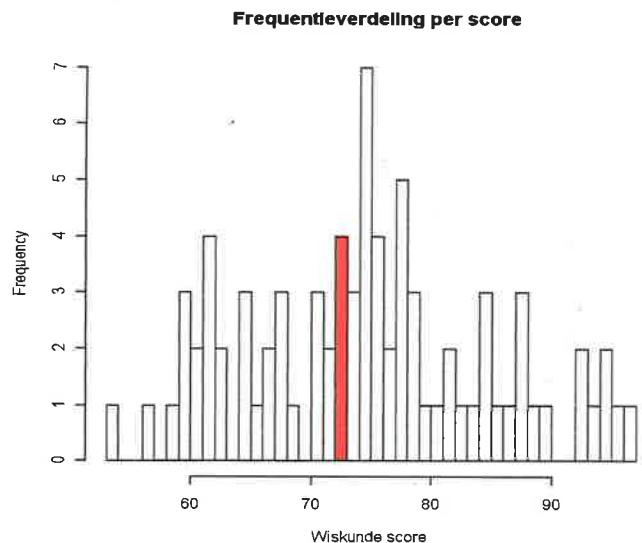


Figuur 8.3: Normaalverdeling voor een variabele met gemiddelde 0 en standaardafwijking 1

We stelden daarnet dat de normaalverdeling een theoretische kansverdeling is. Naast theoretische kansverdelingen bestaan er ook empirische kansverdelingen.

Een **empirische kansverdeling** is gebaseerd op de concrete frequentieverdeling van een variabele. In dat geval zijn het de relatieve frequenties die bepalen wat de kans op de bijbehorende waargenomen waarde is. Naast beschrijven wat er in de steekproef werd waargenomen (cf. empirisch), kan de empirische kansverdeling ook gezien worden als een schatter van de theoretische kansverdeling, die de verdeling beschrijft waaruit de steekproefdata afkomstig zijn. Een relatieve frequentie van een bepaalde waarde wordt dan geïnterpreteerd als de kans dat we in een gelijkaardige theoretische situatie diezelfde waarde zouden waarnemen.

Het voorbeeld van de wiskundescores uit hoofdstuk 5 maakt dit duidelijker. Indien we bij 80 leerlingen wiskundescores meten en vaststellen dat 4 van deze 80 leerlingen een score van 73 behalen, dan bedraagt de relatieve frequentie van de waarde 0,05 (=4/80) (zie ook de frequentietabel op pg 67). Of, de kans dat een willekeurige leerling uit dezelfde populatie in een identieke meetsituatie een score van 73 behaalt is 0,05. Nog anders gesteld, 1 op de 20 willekeurige leerlingen zou een score van 73 behalen.



Figuur 8.4: Frequentieverdeling per wiskundescore met de frequentie voor score 73 rood gemarkeerd

8.1.2

Hieronder geven we een frequentietabel van 20 wiskundescores (Tabel 8.1):



Indien we dezelfde variabele wiskundescores bij gelijkaardige leerlingen in een identieke situatie zouden meten, wat is dan de kans dat een leerling:

- a) een score 84 behaalt?
- b) een score behaalt die hoger is dan 82?
- c) een score behaalt die strikt lager is dan 73?
- d) een score behaalt die niet groter is dan 100?

Tabel 8.1: Frequentie, percentage, absolute cumulatieve frequentie en relatieve cumulatieve frequentie van de wiskundescores van 20 leerlingen

Wiskunde-score	Frequentie (n _i)	Percentage (f _i)	Absolute cumulatieve frequentie	Relatieve cumulatieve frequentie
59,00	1	5,0	1	5,0
60,00	1	5,0	2	10,0
62,00	1	5,0	3	15,0
68,00	2	10,0	5	25,0
71,00	1	5,0	6	30,0
73,00	2	10,0	8	40,0
75,00	2	10,0	10	50,0
76,00	1	5,0	11	55,0
79,00	1	5,0	12	60,0
82,00	1	5,0	13	65,0
84,00	1	5,0	14	70,0
85,00	1	5,0	15	75,0
88,00	2	10,0	17	85,0
90,00	1	5,0	18	90,0
93,00	2	10,0	20	100,0
Totaal (n)	20	100,0		

8.1.3



Naast empirische kansverdelingen (gebaseerd op relatieve frequenties) bestaan er ook theoretische kansverdelingen.

Een **theoretische kansverdeling** is een kansverdeling die gebaseerd is op oneindig veel denkbeeldige kansexperimenten. Zo'n theoretische kansverdeling wordt meestal samengevat (beschreven door) een wiskundige formule of procedure.

Zo wordt de normaalverdeling als volgt in een wiskundige formule samengevat.

$$\text{De kans op waarde } x \text{ voor variabele } X = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Waarbij μ en σ staan voor respectievelijk het gemiddelde en de standaardafwijking van de variabele X .

We hernemen nu dezelfde vraag: waarom is de normaalverdeling zo belangrijk?

Hiertoe kunnen we 2 belangrijke redenen geven:

1. De normaalverdeling is **vaak een goed model voor de verdeling van werkelijke data!**

De verdelingen van de volgende soorten data benaderen vaak de normaalverdeling: scores op een toets die op grote schaal afgenoemt is (bv. wiskundescores, psychologische testen, waarden, ...), het herhaald zorgvuldig meten van eenzelfde kenmerk van biologische populaties (het gewicht van mensen, de lengte van de nek van een volwassen giraf, ...).

Indien variabelen uit onze concrete onderzoekspraktijk (bij benadering) normaal verdeeld zijn, dan kunnen we de normaalverdeling hanteren als model om een aantal kenmerken van de verdeling van onze variabelen af te leiden. Hier gaan we verder in dit OLP dieper op in!

2. De belangrijkste reden is dat de meeste statistische procedures die we zullen hanteren in de inferentiële statistiek gebaseerd zijn op de normale verdeling. Met andere woorden, **de meerderheid van de statistische technieken die we in de inferentiële statistiek zullen hanteren gaan uit van de veronderstelling dat de variabelen waarop we ze willen toepassen normaal verdeeld zijn.**

8.1.4

We hernemen de wiskundescores uit 8.1.2.



Wat is volgens jou de som van de kansen op al deze wiskundescores?

8.1.5

De respons op 8.1.4 brengt ons bij een belangrijk kenmerk van elke kansverdeling en dus ook van de normaalverdeling:



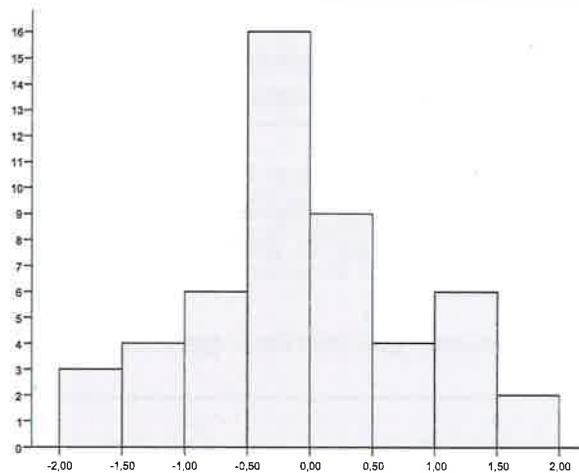
Bij een kansverdeling bedraagt de som van alle kansen 1 (net als de som van alle relatieve frequenties).

Als we dit visualiseren dan wil dat zeggen dat de oppervlakte onder de normaalverdeling exact 1 bedraagt. Of, de blauwe oppervlakte onder de normaalcurve van figuur 8.5 bedraagt exact de waarde 1:



Figuur 8.5: Oppervlakte onder de normaalcurve

Dit wordt duidelijker door eerst te kijken naar een histogram. Neem het onderstaande histogram:



Figuur 8.6: Een voorbeeld van een histogram

In een histogram stellen de oppervlakten van alle staven samen eveneens het totaal aantal waarnemingen voor (bv. 50 in Figuur 8.6). De staven die links van de waarde -0,5 op de X-as liggen hebben een oppervlakte die 13 van de 50 waarnemingen voorstellen. Of die 13/50 (0,26) keer de totale oppervlakte van de staven is. Als we dit histogram hertekenen zodanig dat de Y-as de relatieve frequenties weergeeft, dan zou de totale oppervlakte van de staven de relatieve frequentie van alle waarnemingen samen voorstellen (1). De staven links van de waarde -0,5 zouden samen een oppervlakte hebben die 0,26 van de waarnemingen representeren. De som van alle oppervlakten van de staven zou 1 bedragen.

Analoog stelt de totale oppervlakte onder de normaalcurve de relatieve frequentie (of de kans) voor van alle mogelijke waarden, zijnde 1.

8.1.6



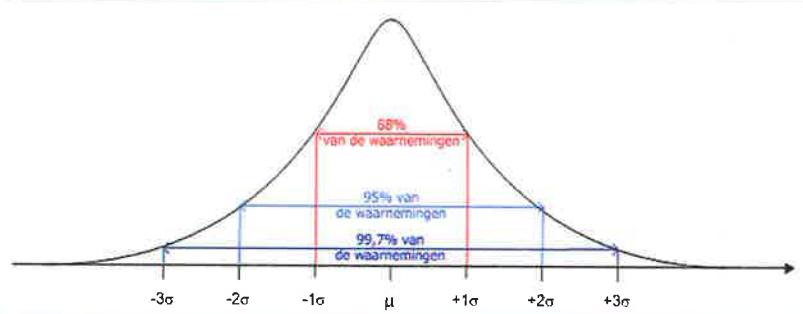
Een belangrijke eigenschap van de normaalverdeling en van een variabele die de normaalverdeling (bij benadering) volgt, is dat wat we de **68-95-99,7-regel** zullen noemen.

Bij de normaalverdeling met een gemiddelde μ en een standaardafwijking σ geldt dat:

- afgerond 68% van de waarnemingen binnen een afstand van -1σ en $+1\sigma$ van het gemiddelde μ bevindt;
- afgerond 95% van de waarnemingen binnen een afstand van -2σ en $+2\sigma$ van het gemiddelde μ bevindt;
- afgerond 99,7% van de waarnemingen binnen een afstand van -3σ en $+3\sigma$ van het gemiddelde μ bevindt.

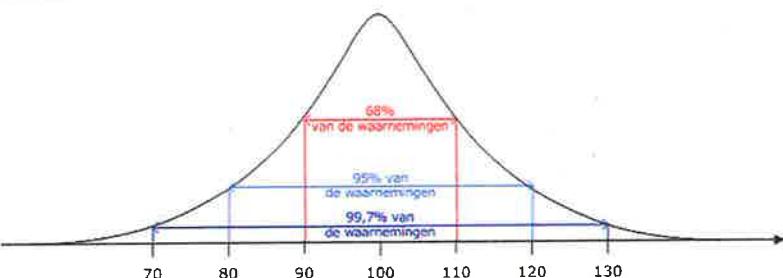
De waarden in bovenstaande kader van één, twee en drie standaardafwijkingen zijn afrondingen. Eigenlijk gaat het om 1 standaardafwijking, 1,96 (i.p.v. 2) en 2,97 (i.p.v. 3). We werken echter meestal met de afgeronde waarden.

Onderstaande figuur stelt deze regel visueel voor:



Figuur 8.7: De 68-95-99,7 regel bij een normale verdeling

Voor een variabele die bij benadering normaal verdeeld is kunnen we een gelijkaardige figuur opstellen. Zo stelt de onderstaande figuur de verdeling voor van een IQ-test met gemiddelde score 100 en een standaardafwijking van 10 punten.



Figuur 8.8: Normaalverdeling voor IQ score

Indien we dus van een variabele weten dat hij de normaalverdeling bij benadering volgt kunnen we via de 68-95-99,7 regel meer accurate informatie afleiden uit de kengetallen gemiddelde en standaardafwijking. Om na te gaan of een variabele de normaalverdeling volgt, baseren we ons op de kengetallen scheefheid en platheid. Dit wordt verder uitgediept in hoofdstuk 10.

Verder zullen we zien dat we bij variabelen die bij benadering normaal verdeeld zijn meer informatie kunnen afleiden door gebruik te maken van een zeer specifieke normaalverdeling: de standaardnormaalverdeling.

8.2. Z-scores

8.2.1



Elke variabele van intervalniveau en hoger wordt uitgedrukt op een specifieke schaal. Zo wordt de lengte van een persoon vaak uitgedrukt in aantal centimeters, de leeftijd in aantal jaren,... Wiskundescores kunnen uitgedrukt worden in een score tussen 0 en 100 of een score op 10, of een score tussen 0 en 200.

Daarnaast zijn er variabelen die op een ‘beteenisloze’ schaal worden uitgedrukt. Dit is vaak het geval bij variabelen die opvattingen of gedragingen kwantificeren. Neem bijvoorbeeld de mate van stuurloosheid van leerlingen in hun leren. Zo’n variabele kan gaan van minimum 1 tot maximum 5 of van minimum 1 tot maximum 30. Ongeacht de schaal die men hanteert, welke betekenis kunnen we dan toekennen aan scores 3,5 of 3,8 ...?

Het is met andere woorden soms nuttig om variabelen uit te drukken op een schaal die meer betekenis heeft en die dezelfde betekenis heeft voor

verschillende variabelen. In wat volgt bekijken we een manier om variabelen uit te drukken op een andere schaal. We spreken dan over **gestandaardiseerde scores** ook wel z-scores genoemd.

Hoe standaardiseer je een variabéle?

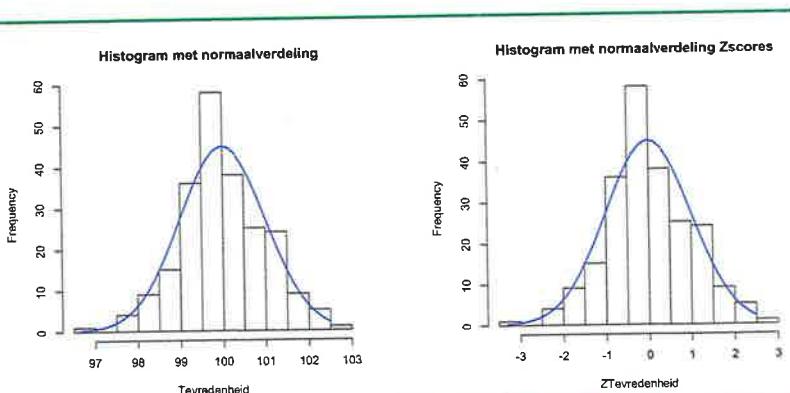
Als x een waarneming is uit een verdeling van een variabele X met een gemiddelde μ en een standaardafwijking σ dan is de gestandaardiseerde score (z) voor deze waarneming:

$$z = \frac{x - \mu}{\sigma}$$

Een gestandaardiseerde score noemen we ook vaak een z-score.

Om een z-score te berekenen, trek je dus eerst de gemiddelde score af van elke waarneming en deel je dit resultaat door de standaardafwijking.

Dit is een lineaire transformatie van de originele schaal waarop de variabele is uitgedrukt. Dit wil zeggen dat deze transformatie de verdeling van de waarnemingen niet beïnvloedt. Dit zie je in de onderstaande twee histogrammen (Figuur 8.9). Histogram 1 (links) geeft de verdeling van een variabele op basis van de originele schaal, histogram 2 (rechts) geeft de verdeling van de z-scores voor deze variabele. Zoals je merkt zijn beide histogrammen identiek!

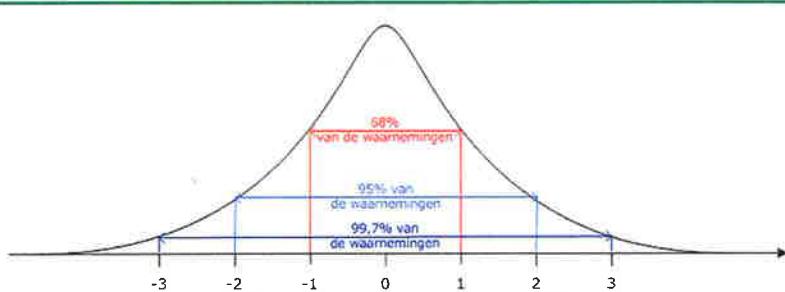


Figuur 8.9: Histogram van tevredenheidsscore en van de z-score voor dezelfde variabele

Hoe kan je de z-scores interpreteren? Z-scores zijn uitgedrukt op een schaal die aangeeft hoeveel standaardafwijkingen een waarneming verwijderd is van het gemiddelde.

Meer concreet: een waarneming met een z-score 0 scoort gemiddeld op het waargenomen kenmerk. Een waarneming met een z-score +1 scoort 1 standaardafwijking hoger dan het gemiddelde. Een waarneming met een z-score -1 scoort 1 standaardafwijking lager dan het gemiddelde. Een waarneming met een z-score van 1,5 scoort anderhalve standaardafwijking hoger dan gemiddeld voor het waargenomen kenmerk. Enzovoort. Dus iemand met een z-score nul op de gestandaardiseerde variabele Tevredenheidsscore die scoort gemiddeld op die variabele. Een persoon met een waarde +1 voor diezelfde gestandaardiseerde variabele scoort één standaardafwijking hoger dan gemiddeld op die variabele.

Bij variabelen die (bij benadering) normaal verdeeld zijn kan je vervolgens op basis van deze z-scores ook de 68-95-99,7-regel toepassen. Hieronder geven we deze regel wederom visueel weer met z-scores:



Figuur 8.10: De 68-95-99,7-regel toegepast op z-scores

68% van de waarnemingen behaalt een z-score tussen -1 en +1, of 16% scoort hoger dan 1 en 16% scoort lager dan -1.

95% van de waarnemingen behaalt een z-score tussen -2 en +2, of 2,5% scoort hoger dan 2 en 2,5% scoort lager dan -2.

99,7% van de waarnemingen behaalt een z-score tussen -3 en +3, of 0,15% scoort hoger dan 3 en 0,15% scoort lager dan -3.

8.2.2 R biedt verschillende mogelijkheden om een variabele te veranderen in een z-score. In onderstaand voorbeeld werken we opnieuw met het bestand 'Opleidingen1'.



MOGELIJKHEID 1:

De z-score kan eenvoudig berekend worden door de wiskundige bewerking toe te passen op de variabele naar keuze. Immers in hoofdstuk 4 leerde je rekenen met variabelen in R.

We passen het toe voor de variabele 'Nut', via het onderstaande commando creëren we een nieuwe variabele Nutz die de z-score bevat:

```
> Opleidingen1$Nutz <- { Opleidingen1$Nut-mean(Opleidingen1$Nut, na.rm=TRUE) }/sd(Opleidingen1$Nut, na.rm=TRUE)
```

MOGELIJKHEID 2

We kunnen dit echter ook eenvoudiger met de eigen aangemaakte functie `zscores()`. Deze vind je terug in het bestand 'OLPfuncties.R'

Na inladen van de functies kan hetzelfde resultaat bekomen worden via

```
> Opleidingen1$Nutz <- zscores(Opleidingen1$Nut)
```

MOGELIJKHEID 3

Ook de functie `scale()` biedt een oplossing. Bij deze functie worden meteen de zscores weergegeven alsook het gemiddelde en de standaardafwijking.

```
> scale(Opleidingen1$Nut, center = TRUE, scale = TRUE)
[1]
[1,] -1.21824860
[2,] -0.06124499
[3,] -1.29797503
[4,] 0.43293082
[5]
[6]
[7]
[265,] 0.83647924
[266,] 0.98038993
[267,] 1.65205835
[268,] 0.03529473
attr(,"scaled:center")
```

```
[1] 0.0416841
attr(,"scaled:scale")
[1] 0.9913512
```

Waarbij het `center=TRUE` staat voor het opvragen van het gemiddelde en `scale=TRUE` de standaardafwijking weergeeft. Echter, de functie `zscores()` is makkelijker om de resulterende z-scores in een nieuwe variabele weg te schrijven.

8.2.3



Hier werken we opnieuw met het databestand Opleidingen1.RData uit hoofdstuk 7. In hoofdstuk 7 leerden we dat de variabele Ndeelnemers het meest normaal verdeeld lijkt.

Beantwoord nu de volgende vragen:

- Welke z-score stemt overeen met 24 deelnemers?
- Welke z-score stemt overeen met 15 deelnemers?
- In de veronderstelling dat deze variabele normaal verdeeld is, wat kan je dan uit de z-scores afleiden over opleidingen met 24 deelnemers?

8.2.4



Het houdt echter niet op bij die ruwe 68-95-99,7 regel. Voor normaal verdeelde variabelen kan uit z-scores veel meer afgeleid worden, op basis van de standaardnormaalverdeling.

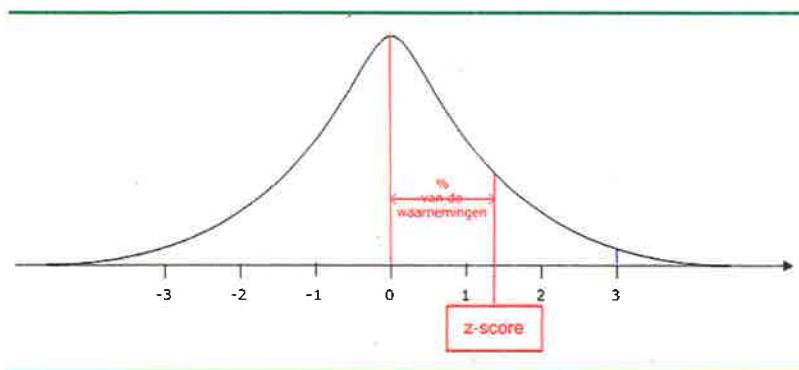
De standaardnormaalverdeling is de theoretische normaalverdeling met gemiddelde 0 en standaardafwijking 1.

Het berekenen van een z-score voor een variabele met een normale verdeling resulteert in een nieuwe variabele die de standaardnormaalverdeling heeft.

Zoals we reeds eerder aanhaalden stellen de oppervlakten onder de normale kromme relatieve frequenties voor.

Voor de standaardnormaalverdeling is door statistici per mogelijke waarde de precieze kans berekend dat deze kan voorkomen. Deze kans staat gelijk aan relatieve frequenties. Deze gegevens zijn door deze statistici opgeliist in tabellen en interactieve software die zo per z-score een bepaalde cumulatieve relatieve frequentie weergeven.

Deze tabel (tabel 8.2) geeft per z-score weer wat de relatieve frequentie is van waarnemingen tussen 0 (het gemiddelde) en de z-score (zie onderstaande figuur).



Figuur 8.11: Illustratie met de ligging van de z-score

Tabel 8.2: Cumulatieve frequenties voor de standaardnormaalverdeling, liggende tussen nul en de z-waarde

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Hoe moet je deze tabel lezen? We illustreren dit aan de hand van enkele voorbeelden op basis van een uitreksel uit de tabel (Tabel 8.3):

Stel dat je wil nagaan hoeveel procent kans je hebt op een z-score tussen 0 en 0,44.

Uiterst links kan je de z-score tot het eerste cijfer na de komma terugvinden (zie rood kadertje rond 0,4).

Bovenaan kan je het tweede cijfer na de komma terugvinden (zie rood kadertje rond 4). Samen identificeert dit de z-score 0,44.

Vervolgens kan je in de tabel zelf de cel zoeken naast 0,4 en onder de 0,04.

In dit voorbeeld gaat het om 0,1700 (zie rood kadertje). Dit houdt in dat 17% van de waarnemingen uit de standaardnormaalverdeling zich bevinden tussen het gemiddelde (nul) en 0,44.

Een ander voorbeeld dat we aanduiden in het groen in de bovenstaande figuur is die van de z-score 1,26. Dit stemt overeen met een oppervlakte tussen 0 en 1,26 die 0,3962 bedraagt. Dit wil zeggen dat 39,62% van de waarnemingen uit de standaardnormaalverdeling zich bevindt tussen 0 en 1,26.

Tabel 8.3: Voorbeeld voor het interpreteren van Z-scores 0,44 en 1,26

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790
1.2	0.3849	0.3869	0.3888	0.3907	0.3927	0.3947	0.3962	0.3980

Voor variabelen die bij benadering normaal verdeeld zijn, kan je aan de hand van z-scores en deze tabel nagaan wat de relatieve frequenties van waarnemingen tussen 0 en bepaalde z-scores zijn. Je kan evenzeer conclusies gaan trekken over de kans dat je bepaalde waarden vaststelt in een nieuwe identieke meetsituatie voor eenzelfde variabele.

Daartoe moet je **rekenen met behulp van deze tabel**. Daarbij moet je ont-houden dat 50% van de waarnemingen lager scoort dan nul en dat de verdeling symmetrisch is. De relatieve frequenties die je kan afleiden voor positieve z-scores gelden ook voor negatieve z-scores.

Zo leert de tabel (zie ook tabel 8.4) ons dat 68% van de waarnemingen een z-score hebben die gelijk is aan 0,47 of lager. Of 32% van de waarnemingen heeft een z-score hoger dan 0,47. Daarnaast leert de tabel ons dat 18% van de waarnemingen een score behaalt tussen -0,47 en 0.

Tabel 8.4: Voorbeeld voor het bepalen van een Z-score a.d.h.v. een tabel

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1551	0.1591	0.1629	0.1661	0.1693	0.1723	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549

8.2.5

We hernemen de oefening op de variabele Ndeelnemers uit Opleidingen1.RData. Bereken indien nodig opnieuw de z-scores. Beantwoord de volgende vragen aan de hand van de z-scores en Tabel 8.2 als je weet dat deze variabele (bij benadering) normaal verdeeld is:

- Hoeveel procent van de opleidingen heeft minder dan 8 deelnemers volgens de normaalverdeling? Klopt dit met onze steekproef?
- Hoeveel procent van de opleidingen heeft 20 of meer deelnemers volgens de normaalverdeling?
- Hoeveel procent van de opleidingen heeft tussen 10 (inclusief) en 15 (exclusief) deelnemers volgens de normaalverdeling?

TIP: bij het rekenen met z-scores is het vaak handig een schets (zoals figuur 8.12) te maken van de normaalverdeling en aan te duiden in welke oppervlakte onder de kromme je geïnteresseerd bent.

8.2.6

Naast het werken met een tabel biedt R ook de mogelijkheid om de berekening van het percentage rechtstreeks te doen met de functie `pnorm()`. Deze functie kan je terugvinden in de standaardversie van R, er moeten geen specifieke pakketten ingeladen worden.



Als voorbeeld hernemen we het bestand Opleidingen1.RData en berekenen we hoe groot de kans is dat er maximum 17 deelnemers zijn. Bij deze functie moeten we het gemiddelde van de variabele Ndeelnemers en de standaardafwijking van dezelfde variabele bovenindien specificeren in respectievelijk het argument `mean=` en `sd=`. Voor deze concrete variabele bedraagt het gemiddelde 15 en de standaardafwijking 4,5.

```
> pnorm(17, mean= 15, sd=4.5)
[1] 0.6716
```

R berekent echter de kans op een andere wijze dan wat we uit tabel 8.2 aflezen. Het resultaat van de `pnorm()` functie geeft het percentage waarnemingen weer dat lager dan of gelijk is aan de opgegeven waarde. Dus, voor de variabele Ndeelnemers stellen we vast dat 67,16% van de opleidingen 17 of minder deelnemers telt.

In dezelfde `pnorm()` functie kunnen we eveneens z-scores invoeren om vervolgens de bijhorende kans op die z-score of lager op te vragen. Daar-toe zouden we bijvoorbeeld het volgende commando kunnen ingeven.

```
> pnorm(1.96, mean=0, sd=1)
[1] 0.9750021
```

Echter, voor het argument `mean=` is de waarde 0 de default waarde in R en voor het argument `sd=` is de waarde 1 de default waarde. Dit maakt dat het onderstaande commando tot hetzelfde resultaat leidt.

```
> pnorm(1.96)
[1] 0.9750021
```

8.2.7

We hernemen nogmaals de oefening op de variabele Ndeelnemers uit Opleidingen1.RData.



Beantwoord de volgende vragen aan de hand van de functie `pnorm`:

- Bij hoeveel procent van de opleidingen kunnen we verwachten dat er strikt minder dan 8 deelnemers zijn?
- Hoeveel procent van de opleidingen behaalt een score hoger dan of gelijk aan 20?

Responsen

Respons 8.1.2

- a) Uit de tabel leren we dat er 1 van de 20 leerlingen een score 84 behaalt. De kans is met andere woorden 1 op 20 of 0,05.
- b) In totaal zijn er 7 leerlingen die een score hoger dan 82 behalen. De kans dat een leerling dus een score hoger dan 82 behaalt bedraagt 0,35. Of 7 op de 20 willekeurige leerlingen zou een score hoger dan 82 behalen.
- c) 6 van de 20 leerlingen behalen een score lager dan 73. De kans dat een leerling lager dan 73 gaat scoren bedraagt met andere woorden 3 op 10 of 0,3.

In deze voorbeelden gebruiken we telkens de relatieve frequenties van een waargenomen variabele om een voorspelling te maken van de kans dat een waarde van dezelfde variabele voorkomt bij een nieuwe meting in een identieke situatie. We gebruiken de huidige waarnemingen om de kans te voorspellen.

- d) De som van alle kansen bedraagt 1 (of 20/20). Dit leidt ons tot een belangrijke eigenschap van een kansverdeling:

Bij een kansverdeling bedraagt de som van alle kansen 1 (net als de som van alle relatieve frequenties).

Respons 8.1.4

De som van alle kansen bedraagt 1 (of 20/20). Dit leidt ons tot een belangrijke eigenschap van een kansverdeling!

Respons 8.2.3

- a) 24 deelnemers stemt overeen met een z-score 1.93.

Dit kon je o.a. berekenen via volgende functie.

```
> (24-mean(Opleidingen1$Ndeelnemers, na.rm=TRUE)) /
sd(Opleidingen1$Ndeelnemers, na.rm=TRUE)
[1] 1.937419
```

- b) 15 deelnemers stemt overeen met een z-score -0.061. Dit komt overeen met ongeveer 0. We kunnen dus stellen dat 15 deelnemers in de buurt komt van het gemiddelde aantal deelnemers.

```
> (15-mean(Opleidingen1$Ndeelnemers, na.rm=TRUE)) /
sd(Opleidingen1$Ndeelnemers, na.rm=TRUE)
[1] -0.06132122
```

- c) Een opleiding met 24 deelnemers scoort ongeveer twee standaardafwijkingen hoger dan het gemiddelde aantal. We kunnen dus stellen dat een opleiding met 24 deelnemers tot de 2.5% meest gevulde opleidingen hoort.

Respons 8.2.5

- a) Hoeveel procent van de opleidingen heeft minder dan 8 deelnemers?

Om deze vraag te beantwoorden moeten we eerst de z-score berekenen voor de waarde 8. Deze z-score bedraagt -1.61589703 afggerond -1.62.

Vervolgens gaan we in de tabel na welke cumulatieve relatieve frequentie overeenstemt met deze z-score. Daarom kijken we in de tabel onder de z-score 1.62 (aangezien er geen negatieve z-scores zijn opgenomen en de verdeling symmetrisch is).

Voor deze waarde lezen we uit de tabel het volgende cijfer af: 0.4474. Met andere woorden 44.74% van de leerlingen behaalt een z-score tussen 0 en -1.62

Daarnaast weten we dat 50% van de leerlingen een score lager dan gemiddeld behaalt (=eigenschap van de (standaard)normale verdeling). Of 50% behaalt een z-score lager dan nul.

Om te weten te komen hoeveel procent van de opleidingen minder dan 8 deelnemers heeft en dus een z-score -1.62 dienen we 44.74% af te trekken van 50%. Met andere woorden 5.26%. We kunnen dit ook in kansen uitdrukken: de kans dat een willekeurige opleiding minder dan 8 deelnemers heeft is 5.26%.

Klopt dit met onze steekproef?

Hiervoor maken we opnieuw de frequentietabel. Hierin zien we dat 15 van de 268 opleidingen 8 of minder deelnemers hadden. Dit komt overeen met 5.6%. Dit komt dus redelijk goed overeen.

freqtabel (Ndeelnemers)

X	Freq	Percentage	CumulativeN	CumulativePerc
1	4	2	2	0.7462687
2	5	1	3	0.3731343
3	6	2	5	0.7462687
4	7	6	11	2.2388060
5	8	4	15	4.1044776
6	9	9	24	5.5970149
7	10	16	40	8.9552239
8	11	23	63	14.9253731
9	12	12	75	23.5074627
10	13	16	91	27.9850746
11	14	19	110	33.9552239
12	15	35	145	41.0447761
13	16	23	168	54.1044776
14	17	19	187	62.6865672
15	18	20	207	69.7761194
16	19	15	222	77.2388060
17	20	10	232	82.8358209
18	21	9	241	86.5671642
19	22	9	250	89.9253731
20	23	5	255	93.2835821
21	24	8	263	95.1492537
22	25	3	266	98.1343284
23	26	1	267	99.2537313
24	27	1	268	99.6268657
				100.0000000

b) Hoeveel procent van de opleidingen heeft meer dan 20 deelnemers?

Hier is de redenering identiek. Echter, we moeten de z-score voor de waarde 19 opzoeken. Enkel zo komen we te weten hoeveel procent van de opleidingen tussen 0 en 19 deelnemers had. De z-score voor 19 deelnemers bedraagt 0,83. Uit de tabel lezen we af dat 29,67% van de waarnemingen uit de standaardnormaalverdeling ligt tussen de waarde nul en 0,83. In totaal scoort 50% hoger dan gemiddeld. Om te weten hoeveel procent van de waarnemingen uit de standaardnormaalverdeling boven 0,83 ligt volstaat het om 29,67 af te trekken van 50 (=20,33%). 20,33% van de waarnemingen uit de standaardnormaalverdeling ligt boven de z-score 0,83. Aangezien 19 deelnemers overeenstemt met de een z-score 0,83 kunnen we ervan uitgaan dat 20,33% van de opleidingen 20 of meer deelnemers telde.

c) Hoeveel procent heeft tussen de 10 en 15 deelnemers?

Hiervoor berekenen we eerst z-scores voor beide aantalen:

- z-score voor 10=-1.1717 afgerond -1.17;
- z-score voor 15=-0,0613 afgerond -0,06;

Daarna zoeken we de proporties op in de tabel:

- voor z-score -1,17 is dit 0,3790;
- voor z-score -0,06 is dit 0,0239;

Daarna berekenen we de proportie waarnemingen in de standaardnormaalverdeling met een z-score -1,17 tot en met -0,06 door de volgende formule:

$$-0,3790-0,0239=0,3551;$$

Of, 35,51% van de waarnemingen in een standaardnormaalverdeling situeert zich tussen een z-score -1,17 en -0,06.

Vertaald naar ons aantal deelnemers wil dit zeggen dat bij benadering 35,5% van de opleidingen tussen de 10 en de 15 deelnemers zit.

Respons 8.2.7

a) Om een beeld te krijgen op het percentage opleidingen met minder dan 8 opleidingen dienen we de cumulatieve propoertie op te zoeken van de waarde 7 in de normaalverdeling met een gemiddelde van 15,28 en een standaardafwijking van 4,5:

```
> pnorm(7, mean=15.28, sd=4.5)
[1] 0.03288412
```

3,29% van de opleidingen heeft dus tussen de 0 en de 7 deelnemers, of 3,29% van de opleidingen heeft minder dan 8 deelnemers.

b) Om te weten te komen hoeveel procent van de opleidingen 20 of meer deelnemers had, zoeken we de cumulatieve relatieve frequentie op van de waarde 19 in de normaalverdeling met gemiddelde 15,28 en standaardafwijking 4,5.

```
> pnorm(19, mean=15.28, sd=4.5)
[1] 0.795787
```

79,58% van de opleidingen heeft dus tussen de 0 en de 19 deelnemers.

Wanneer we willen weten hoeveel opleidingen meer dan 20 deelnemers hebben trekken we dit getal af van 100% en komen zo tot 20,42% wat nagenoeg overeenkomt met het antwoord uit 8.2.5 b), op afrondingsfoutjes na.

Gehanteerde functies

Functie	Doelstelling	Bron
<code>pnorm()</code>	Bij deze functie wordt rechtstreeks het percentage bereken voor een waarde x. In deze functie dienen eveneens de argumenten <code>mean=</code> en <code>sd=</code> te worden ingegeven. vb: <code>pnorm(13, mean= 17, sd=2.5)</code>	R basispakket
<code>scale()</code>	Bij deze functie worden meteen de zscores weergegeven alsook het gemiddelde en de standaardafwijking. Waarbij het <code>center=TRUE</code> staat voor het opvragen van het gemiddelde en <code>scale=TRUE</code> de standaardafwijking weergeeft.	R basispakket
<code>zscore()</code>	Bij deze functie worden meteen de zscores weergegeven. Deze functie is handig om meteen een nieuwe variabele aan te maken.	OLP functies.R

HOOFDSTUK 9

Steekproeftheorie

DOELSTELLINGEN:

Na dit hoofdstuk

- weet je wat een populatie en een steekproef zijn;
- weet je wat bedoeld wordt met de Wet van de Grote Getallen;
- weet je wat het onderscheid is tussen een steekproefverdeling en de steekproevenverdeling;
- weet je wat de standaardfout is;
- ken je de toepassing van de Centrale Limietstelling;
- kan je een steekproef uit een kolom in R trekken.

NODIGE FILES:

Snoepjes.RData

OLP Functies.R

een file met daarin aangepaste functies die bij dit OLP horen



Statistiek kan je helpen om waargenomen gegevens te beschrijven. Later zullen we zien dat er een beroep wordt gedaan op statistiek om sociale fenomenen op basis van gegevens te verklaren en zelfs voorspellingen te formuleren.

Deze beschrijvende, verklarende of voorspellende bezigheden zijn zowel bij populatieonderzoek (waarbij we alle eenheden waarover we een uitspraak willen doen betrekken in het onderzoek) als bij steekproefonderzoek (waarbij we een selectie van de eenheden waarover we een uitspraak willen doen, betrekken) noodzakelijk. Echter, populatieonderzoek is eerder uitzondering dan regel omdat we vaak in onderzoek niet alle observatie-eenheden kunnen betrekken. Doorgaans maken we gebruik van steekproeven. Een vraag die dit automatisch oproept, luidt: kunnen we de conclusies op basis van de steekproefgegevens ook veralgemenen naar de populatie toe? Het is precies deze vraag die de kern uitmaakt van **wat we inferentiële statistiek noemen**. In dit hoofdstuk staan we stil bij de kernbegrippen uit de steekproeftheorie die ons zullen leiden om gefundeerde uitspraken te doen over de populatie op basis van steekproefgegevens.

9.1. Wat is een populatie?

9.1.1



Stel dat je wilt weten wat de houding is van leerlingen uit het secundair onderwijs ten aanzien van geweld op school. Na het formuleren van een probleemstelling (wat is de houding van leerlingen ten aanzien van geweld) en het uitwerken van een onderzoeksplan heb je een vragenlijst ontwikkeld met vragen omtrent de opvattingen ten aanzien van antisociaal en crimineel gedrag.

We gaan er van uit dat je de juiste vragen hebt geformuleerd en de goede instrumenten hebt ontwikkeld. Dan nog zit je met een belangrijke vraag.

Aan wie zal je die vraag voorleggen? Wat zou jij doen?

9.1.2



Het bovenstaande probleem toont dat je bij elk onderzoek een aantal belangrijke keuzes moet maken. Het welslagen van je onderzoek hangt grotendeels af van de keuzes die je hierin maakt.

Ga je iedereen bevragen of beperk je je tot een selectie uit die populatie?

Hoewel je in het bovenstaande probleem geïnteresseerd bent in de verdeling van opvattingen bij al de scholieren uit het secundair onderwijs, zal het vaak niet mogelijk zijn om ze allemaal te bevragen.

Een eerste stap bij het maken van een keuze over wie je bevraagt, is het bepalen van de populatie.

Onder **populatie** verstaan we doorgaans **de verzameling eenheden** waarover we in ons onderzoek een uitspraak willen doen.

9.1.3



- a) Wat is de populatie in het geschetste onderzoeksprobleem uit 9.1.1 ?
- b) Vind je dit een werkzame omschrijving van de populatie? Kan je er met andere woorden meteen uit afleiden wie je zou moeten bevragen?

9.1.4



Een zeer precieze bepaling van de populatie vergemakkelijkt de keuze van de te bevragen respondenten. Vandaar dat ervoor gepleit wordt om de populatie uit je onderzoek zo precies mogelijk te omschrijven. Dit behoedt je ook voor ongeoorloofde uitspraken. Zo zouden conclusies aangaande de waarden van leerlingen uit het bso en tso uit het huidig schooljaar, op basis van een bevraging van leerlingen uit het aso van schooljaar 1997-1998 geen valide conclusies zijn:

Voorbeelden van meer precieze omschrijvingen van populaties zijn:

- de populatie van 50 doelen van een voetballer;
- de populatie bestaande uit alle UA studenten van het academiejaar 2010-2011;
- de populatie bestaande uit alle Vlaamse gezinnen;
- de populatie bestaande uit alle georganiseerde opleidingen in een Centrum voor Volwassenenonderwijs;
- de populatie bestaande uit alle Belgische gemeenten.

9.1.5



In de statistiek maakt men doorgaans ook het onderscheid tussen **eindige** en **oneindige populaties**.

Eindige populaties zijn bestaande verzamelingen, zoals de populatie van alle leerlingen 2de en 3de graad aso in Vlaanderen, van het schooljaar 2009-2010. Het zijn dus populaties die begrensd zijn in hun omvang.

Oneindige populaties zijn theoretische uitvindingen. Het gaat om constructies, zoals alle mogelijke resultaten (kop / munt) van het opeenvolgend tossen van een muntstuk. Je kan daar mee blijven doorgaan, en als je dood valt, kan de taak door iemand anders worden overgenomen, *ad infinitum*.

9.2. Steekproeven

9.2.1



Meestal ben je wel geïnteresseerd in de eigenschappen van de gehele populatie, zoals bijvoorbeeld in de opvatting van leerlingen uit de 2de en 3de graad secundair onderwijs in Vlaanderen, ten aanzien van geweld op school. Maar het is vaak niet mogelijk alle eenheden van die populatie op te nemen in het onderzoek. Dat kan het geval zijn omdat die populatie heel verspreid is, heel omvangrijk is,... Het kan te duur of een te lang durende onderneming zijn om alle eenheden te bereiken.

Meestal zal men toevlucht nemen tot het bestuderen van **een gedeelte van de populatie: een steekproef**.

Een steekproef is dus een uitweg in onderzoek, geboren uit de noodzaak om tijd, energie en middelen te besparen en toch iets belangwekkend te weten te komen van een populatie.

9.2.2



Stel dat je geïnteresseerd bent in de verdeling van de kleuren van de ballen uit het ballenbad van Ikea. Je weet dat er gele, groene, blauwe en rode balletjes zijn en je weet dat er meer dan 5000 balletjes in het bad zitten.

Hoe zou je intuïtief eraan beginnen om de verhoudingen van elke kleur te schatten?

9.2.3



Analoog ga je te werk bij het bevragen van de leerlingen uit het secundair onderwijs:

- je gaat toevallig een aantal scholen kiezen,
- je gaat er leerlingen toevallig kiezen uit de 2de en 3de graad

Het motto bij steekproeftrekking luidt altijd: hoe meer, hoe beter. Later plaatsen we daar wel enkele kanttekeningen bij.

We ondervragen een beperkt aantal leerlingen in de hoop dat dit staal van leerlingen ons een waarheidsgetrouw beeld levert van het geheel. Maar

we kunnen ons vergissen. We kunnen een "goede" (lees waarheidsgetrouwe) steekproef trekken, maar we kunnen ook een "slechte" (lees een vertekende) steekproef eruit halen.

9.2.4 Stel dat je wilt weten hoe de Vlaamse bevolking zou reageren op de aanstelling van een vrouw als minister-president van de Vlaamse regering.

Je kan dan alle stemgerechtigden bevragen (van 18 jaar tot ...), maar dat is onhaalbaar. Een alternatief is het trekken van een steekproef van de mensen die de meeste kans hebben om te gaan stemmen: 18 tot en met 75 jarigen met de Belgische nationaliteit.

Onder het motto "hoe meer, hoe beter" heb je de middelen verkregen om 10000 mensen te bevragen. Stel dat je ze toevallig kiest, maar dat je toevallig 9950 mannen bevraagt en 50 vrouwen. Is dat een goede steekproef? Waarom wel of niet?

9.2.5 Zodra je met steekproefgegevens werkt, ben je geïnteresseerd in welke mate het beeld dat de steekproef oplevert een goed beeld is voor de gehele populatie. Je wilt bijvoorbeeld weten hoeveel mensen uit je steekproef instemmen met de stelling dat een vrouwelijke minister-president even geschikt is. Dit kan bijvoorbeeld na je analyse leiden tot een uitspraak als "7398 respondenten is van mening dat een vrouw niet even bekwaam is om minister-president te zijn van de Vlaamse regering als een man. 2602 respondenten vinden dat een vrouw even bekwaam is als een man om minister-president te zijn van de Vlaamse regering."

Op zich is die 7398 wel interessant, maar je bent daarenboven geïnteresseerd om te weten wat de verhouding is van dit aandeel op het geheel. Dus 7398 op 10000, ofwel 74%.

Hierbij blijft het echter niet. Je wilt deze informatie gebruiken om iets meer te vertellen over de populatie waaruit je je eenheden hebt gekozen. Je wilt conclusies trekken voor een populatie op basis van een analyse van een steekproef. Hierbij vergelijken statistici een waargenomen verdeling (7398-2602) tegenover een theoretische verdeling (niet opgebouwd uit waarnemingen, maar op basis van een redenering).

Deze hele **vergelijkende** fase, waarbij wordt rekening gehouden met de condities van de steekproef en met alle mogelijke theoretische steekproe-

ven die je je maar kan inbeelden, wordt ook wel de **inductieve of inferentiële statistiek** genoemd. Deze vorm van statistiek formuleert met andere woorden uitspraken over de mate waarin onze beweringen (bijvoorbeeld gemiddelden of percentages met een bepaald kenmerk, of bepaalde opvatting) kunnen opgaan voor de gehele populatie.

Aangezien we nooit met zekerheid kunnen stellen dat een steekproefresultaat ook geldig is voor de ganse populatie, werken we bij inferentiële statistiek steeds met een welbepaald betrouwbaarheidsniveau of een welbepaald significantieniveau. Bijvoorbeeld een betrouwbaarheidsniveau van 95% geeft aan dat we ons in 5% van de gevallen zullen vergissen in onze uitspraak over de gehele populatie. We komen hier later op terug.

9.3. De ene steek is de andere niet

9.3.1 Het mag duidelijk zijn dat we continu worden bestookt met cijfers die berusten op steekproefgegevens. Ook gegevens waarmee sociale wetenschappers werken, hebben zelden betrekking op een hele populatie. De meeste onderzoekers behelpen zich met een steekproef uit die populatie. Maar steekproeven omvatten steeds een zekere marge van onnauwkeurigheid. We gaan daar dadelijk verder op in. Maar eerst leren we je in R een willekeurige steekproef van waarden trekken uit een bepaalde kolom (variabele) uit een dataset. We gaan doorheen dit hoofdstuk geregeld gebruik maken van die functie. Nadat we dit uitstapje in R gemaakt hebben, verdiepen we ons verder in de steekproeftheorie aan de hand van een uitgewerkt voorbeeld in R.

9.3.2 Doorheen dit hoofdstuk zullen we geregeld vragen om een steekproef van waarnemingen te nemen uit een variabele. Daartoe kunnen we gebruik maken van de `sample()` functie in R. Tussen de haakjes verwijst je eerst naar de kolom waaruit je een steekproef van waarden wilt nemen. Vervolgens laat je dat volgen door een komma en dan de steekproefgrootte die je wenst. Toegepast geeft dit:

```
sample(Dataset$Variabele, n)
```

`Dataset$Variabele` is daarbij de verwijzing naar de kolom waaruit een waarden getrokken moeten worden. `n` verwijst naar de omvang van de steekproef.

Laten we dit even illustreren in R. We maken eerst een kolom aan die we de naam Waarden geven, met daarin tien cijfers tussen één en tien.

```
Waarden<-rep(1:10)
```

Laat ons nu een steekproef nemen van drie getallen. Dit doen we als volgt. Let daarbij op dat we hier niet gebruik maken van een kolom uit een dataset en bijgevolg ook niet moeten verwijzen naar een dataset noch dat we het \$-teken nodig hebben.

```
sample(Waarden,3)
[1] 6 4 9
```

Belangrijk om te weten is dat telkens je dit commando opnieuw geeft, het resultaat anders kan zijn. Je kan zowel hier als in de toepassingen later in dit hoofdstuk bijgevolg andere resultaten bekomen. De wijze waarop we nu een steekproef getrokken hebben, is bovendien wat heet "zonder teruglegging": de getallen 6, 4 en 9 bv. worden niet terug in de poel van getallen geplaatst. We komen bijgevolg altijd drie verschillende getallen uit. De waarde twee zal geen tweemaal in de steekproef voorkomen. Je kan eveneens een steekproef trekken "met teruglegging". Dit houdt in dat je na het trekken van het eerste getal opnieuw een steekproef uit alle mogelijke getallen tussen één en tien neemt inclusief het eerst getrokken getal. Echter, deze methode heb je in sociaal wetenschappelijk onderzoek zelden nodig.

9.3.3

De kinderen van Zichen-Zussen-Bolder (ZZB) gaan de laatste jaren met Halloween op zoek naar snoep in de buurt. Ze bellen bij iedere inwoner aan en verzamelen al het lekkers dat ze kunnen krijgen. De 100 kinderen die het gehucht rijk is, hebben dit jaar een aanzienlijke buit binnengerijfd. Hieronder staat voor elk achtjarig kind van ZZB aangegeven hoeveel stukken snoep het buit kunnen maken.

33	42	28	43	34	48	31	45	40	47
37	35	39	37	30	27	33	30	23	18
31	33	38	44	37	31	48	39	32	36
33	14	41	8	47	39	35	37	34	30
8	10	11	20	23	17	22	21	28	7
48	39	45	36	38	48	40	40	42	46
25	23	33	17	37	43	45	33	9	44
33	40	37	29	44	39	31	38	27	37
24	34	37	31	36	32	38	38	45	29
21	14	7	5	18	2	9	11	7	18

De populatie bestaat in dit geval uit 100 kinderen. Deze data staan ook in het databestand Snoepjes.RData.

- Stel dat je wilt weten hoeveel stukken snoep elk kind over het algemeen heeft verzameld, welke parameter van ligging zou je dan gebruiken om dit het best te voorspellen? Bereken de relevante parameter in R.
- Selecteer nu een toevallige steekproef van twee kinderen. Maak daarbij gebruik van het `sample()` commando en plaats het resultaat in een object dat je Steekproef1 noemt. Bereken voor deze steekproef de relevante parameter van ligging in R en schrijf het resultaat weg in een object dat je de naam Steek1 noemt.
- Herhaal dit proces nog negen keer. Bereken telkens de relevante parameter van ligging voor elk van die negen steekproeven afzonderlijk. Schrijf het resultaat telkens weg in een object dat je de naam Steek2 tot Steek10 noemt.
- Selecteer nu twee nieuwe willekeurige steekproeven met een omvang van tien kinderen. Bereken opnieuw het gemiddelde voor elk van die steekproeven en schrijf het resultaat weg in respectievelijk Steek11 en Steek12.
- Tot slot herhaal je dit proces voor twee steekproeven met een steekproefomvang van 50. De gemiddelden die deze twee steekproeven opleveren schrijf je weg in respectievelijk Steek13 en Steek14.
- Wat is je conclusie als je al de gemiddeldes uit elk van die verschillende steekproeven naast elkaar zet?

9.3.4

Dit brengt ons bij een ijzeren wet uit de steekproeftheorie:

De wet van de grote getallen:

Hoe groter de steekproefomvang, hoe "juister" het beeld dat de steekproef oplevert van de populatie en dus hoe dichter het steekproefgemiddelde bij het populatiegemiddelde zal liggen.

Dit kunnen we illustreren aan de hand van het volgende voorbeeld. Stel dat we geïnteresseerd zijn om te weten hoe de leeftijd verdeeld is in de Belgische bevolking volgens sekse. In dat geval zou het ons ontzettend

veel energie kosten om de hele bevolking via een enquête te bevragen. De Belgische overheid heeft dat voor ons gedaan in 2001. Toen werd namelijk informatie verzameld via de Socio-Economische Enquête 2001 over alle levende personen die op het Belgisch grondgebied waren geregistreerd.

De resultaten van de opdeling naar leeftijd en sekse vind je in onderstaande tabel.

Tabel 9.1: Absoluut aantal mannen en vrouwen per leeftijdscategorie op basis van de Socio-Economische Enquête 2001

	Populatie		Rijtotaal
	Mannen	Vrouwen	
- 14 jaar	922521	882300	1804821
15-19	308617	295208	603825
20-24	322683	316372	639055
25-29	336624	329303	665927
30-34	377550	366903	744453
35-39	411046	399610	810656
40-44	399036	391593	790629
45-49	369718	364162	733880
50-54	347134	342524	689658
55-59	280865	283407	564272
60-64	244574	260256	504830
65-69	235248	267777	503025
70-74	205423	262199	467622
75+	274407	499290	773697
Kolomtotaal	5035446	5260904	N=10296350

Het totaal aantal geregistreerde mensen op het Belgisch grondgebied bedroeg toen 10 296 350. De gemiddelde leeftijd voor mannen bedroeg 39,1 jaar en voor vrouwen 41,6 jaren.

Als we deze aantallen omzetten naar totaalpercentages bekomen we de volgende tabel:

Tabel 9.2: Relatief aantal mannen en vrouwen per leeftijdscategorie op basis van de Socio-Economische Enquête 2001

	Populatie		Rijtotaal
	Mannen	Vrouwen	
- 14 jaar	9,0%	8,6%	17,5%
15-19	3,0%	2,9%	5,9%
20-24	3,1%	3,1%	6,2%
25-29	3,3%	3,2%	6,5%
30-34	3,7%	3,6%	7,2%
35-39	4,0%	3,9%	7,9%
40-44	3,9%	3,8%	7,7%
45-49	3,6%	3,5%	7,1%
50-54	3,4%	3,3%	6,7%
55-59	2,7%	2,8%	5,5%
60-64	2,4%	2,5%	4,9%
65-69	2,3%	2,6%	4,9%
70-74	2,0%	2,5%	4,5%
75+	2,7%	4,8%	7,5%
Kolomtotaal	48,9%	51,1%	100,0%

De volgende drie tabellen geven een overzicht van drie willekeurige steekproeven uit diezelfde populatie. De eerste steekproef bestaat uit 30 personen (tabel 9.3), de tweede bestaat uit 300 personen (tabel 9.4) en de derde omvat 3000 personen (tabel 9.5).

Tabel 9.3: Relatief aantal mannen en vrouwen per leeftijdscategorie op basis van een steekproef van 30 willekeurige personen uit de Socio-Economische Enquête 2001

	Geslacht		Rijtotaal
	Man	Vrouw	
17-25	17,4%	4,3%	17,4%
25-35	13,0%	17,4%	17,4%
35-45	17,4%	8,7%	34,8%
45-55	4,3%	4,3%	8,7%
55-65	8,7%		8,7%
65-75	4,3%		8,7%
75+			4,3%
Kolomtotaal	65,2%	34,8%	100,0%

We zien onmiddellijk dat de percentages vaak hard afwijken van de percentages in de populatie. Dat deze steekproef afwijkende gegevens oplevert, blijkt ook uit de gemiddeldes: de gemiddelde leeftijd voor de mannen bedraagt 32,7 jaar en de gemiddelde leeftijd van de vrouwen in de steekproef bedraagt 33,7 jaar als we ons baseren op de steekproef met grootte 30.

Tabel 9.4: Relatief aantal mannen en vrouwen per leeftijdscategorie op basis van een steekproef van 300 willekeurige personen uit de Socio-Economische Enquête 2001

	Geslacht		Rijtotaal
	Man	Vrouw	
17-25	9,0%	8,6%	17,6%
25-35	4,1%	5,6%	9,7%
35-45	10,1%	7,9%	18,0%
45-55	11,2%	9,7%	21,0%
55-65	7,1%	5,2%	12,4%
65-75	6,0%	7,9%	13,9%
75+	2,6%	4,9%	7,5%
Kolomtotaal	50,2%	49,8%	100,0%

Volgens de steekproef van 300 Belgen zou de gemiddelde leeftijd voor de mannen 41,5 jaar zijn, terwijl de gemiddelde leeftijd van de vrouwen op 41,7 jaar komt. Dit komt al aardiger in de buurt van de populatiegemiddelen. De percentages uit tabel 9.4 zijn eveneens minder afwijkend van de populatiepercentages.

Tabel 9.5: Relatief aantal mannen en vrouwen per leeftijdscategorie op basis van een steekproef van 3000 willekeurige personen uit de Socio-Economische Enquête 2001

	Geslacht		Rijtotaal
	Man	Vrouw	
17-25	9,8%	9,4%	19,2%
25-35	7,4%	8,4%	15,8%
35-45	8,7%	8,7%	17,4%
45-55	8,2%	8,0%	16,2%
55-65	5,5%	6,5%	12,0%
65-75	5,2%	5,8%	11,0%
75+	3,1%	5,4%	8,5%
Kolomtotaal	47,8%	52,2%	100,0%

In de laatste steekproef (met grootte 3000) hebben de mannen een gemiddelde leeftijd van 38,4 terwijl de vrouwen gemiddeld 41,0 jaar oud zijn. Deze gemiddeldes leunen nog dichter aan bij de populatiegemiddelden. Hetzelfde kan ook gezegd worden van de percentages. Uit deze illustratie blijkt duidelijk de invloed van de steekproefgrootte. Hoe groter de steekproef, hoe preciezer de schattingen.

9.4. Fouten in steekproeven

9.4.1



Het is duidelijk dat elke steekproef uit een populatie een zekere on nauwkeurigheid bevat. Telkens we een steekproef trekken uit een populatie, zal er steeds een zekere mate van onzekerheid zijn betreffende de mate waarin onze steekproef een goede afspiegeling is van die populatie. Hiermee komen we tot de term **representativiteit**.

We zijn steeds in het ongewisse wat betreft de representativiteit van onze steekproef. Zijn de drie steekproeven uit de Sociale Enquête (zie 9.3.3) goede afspiegelingen van de populatie of niet? Hier kunnen we dat inschatten omdat we de populatie kennen. Maar wat zijn dan de criteria om te kunnen spreken van een “goede afspiegeling”?

Als we een parameter berekenen op basis van een steekproef, kunnen we eigenlijk nooit zeker zijn dat we een goede weergave bieden van de populatieparameter. Elke steekproef impliceert bijgevolg een **steekproeffout**. Doordat we niet alle eenheden van de populatie in ogenschouw nemen bij het berekenen van zo’n parameter als het populatiegemiddelde, maken we mogelijk een fout. Stel dat we bijvoorbeeld de gemiddelde leeftijd willen kennen van de mannen, dan kunnen we op basis van de getrokken steekproeven nooit zeker zijn of de berekende gemiddelden goede schattingen zijn voor de gemiddelde leeftijd in de populatie Belgische mannen. De mate waarin de gemiddelde leeftijd van mannen van de eerste steekproef met 30 respondenten (32,7 jaar) afwijkt van de relevante populatieparameter, de gemiddelde leeftijd van de Belgische man (39,1 jaar), noemen we de **steekproeffout**.

In het geval van een steekproef van 30 uit de Belgische bevolking kan deze afwijking vrij groot zijn. Stel dat we een steekproef van 90% uit die bevolking zouden nemen, dan zouden we een betere schatting krijgen van de gemiddelde leeftijd (zie: **de Wet van de grote getallen**). Toch zal er steeds een verschil bestaan tussen de geschatte gemiddelde leeftijd en de werkelijke gemiddelde leeftijd. Aangezien we niet de leeftijdgegevens hebben van alle Belgen zullen we het populatiegemiddelde steeds over- of unterschatten op basis van steekproefgegevens.

Als we een groot aantal eenheden **toevallig** kiezen, hebben we meer kans om mannen te kiezen van beide uiteinden van de verdeling. Anders gezegd, we hebben dan meer kans om zowel jonge, als oude mannen te kiezen. Dit betekent niets meer of minder dan dat, als je steekproef-

grootte toeneemt, ook de kans toeneemt dat je de diversiteit uit de populatie ook *reproduceert* in je steekproef. Dit betekent dat je steekproef meer kans heeft om representatief te zijn. En dit betekent dat je gewoon meer kans hebt dat het geschatte gemiddelde in de steekproef een goede *benadering* vormt van het populatiegemiddelde.

We moeten als onderzoekers er ons bewust van zijn dat het bij elk kengetal dat we afleiden op basis van een steekproef in feite maar om een schatting gaat van het kengetal voor de hele populatie. En dit gaat op voor elk kengetal (of ruimer gesteld parameter): dus zowel voor kengetallen voor ligging, spreiding als vorm. Later zullen we dit nog verder veralgemenen voor parameters die ook samenhang tussen variabelen getalmatig uitdrukken. Voor elk kengetal is er dus sprake van een steekproeffout.

Daarnaast moeten we ons ervan bewust zijn dat de wet van de grote getallen bestaat: naarmate je steekproef groter is, zal de relevante steekproefschatting (zoals een gemiddelde, maar ook een standaarddeviatie, scheefheid, kurtosis, enz.) een kleinere steekproeffout vertonen, en bijgevolg een nauwkeurigere schatting opleveren van de overeenstemmende populatieparameter.

9.4.2



Herneem het voorbeeld van het snoep van de Halloween-kinderen van Zichen-Zussen-Bolder (9.3.3).

a) Eerder nam je tien verschillende steekproeven van telkens twee eenheden. Het gemiddeld aantal snoepjes voor elk van die steekproeven schreef je weg in de objecten Steek1-Steek10. Maak een nieuwe kolom aan met de naam Gemiddelen1, waarin je al deze gemiddelden uit Steek1 t.e.m. Steek10 wegschrijft. Daartoe kan je het volgende commando hanteren:

```
> Gemiddelen1<-c(Steek1, Steek2, Steek3, Steek4, Steek5,
  Steek6, Steek7, Steek8, Steek9, Steek10)
```

Neem nu het gemiddelde van die nieuwe "variabele" Gemiddelen1 en schrijf daarvan het resultaat weg in Schatting1.

b) Bereken nu voor de gemiddelden op basis van de twee steekproeven van tien kinderen (Steek11 en Steek12) ook het gemiddelde. Schrijf het gemiddelde van beide gemiddelden weg in Schatting2.

c) Vergelijk zowel Schatting1 als Schatting2 met het werkelijke populatiegemiddelde. Wat kan je hieruit concluderen?

9.4.3



Als je oefening 9.4.2 hebt uitgevoerd dan kan je niet om de vaststelling heen dat het gemiddelde van verschillende steekproefgemiddelen een betere benadering geeft van het populatiegemiddelde dan de individuele steekproefgemiddelen. Daarnaast kan je vaststellen dat deze schatting precieser wordt – dus dat het geschatte gemiddelde dichter in de buurt van het werkelijke gemiddelde komt – naarmate de steekproefgrootte toeneemt.

Beide vaststellingen zijn de twee eerste kernelementen van een cruciaal onderdeel in de steekproeftheorie: de zogenaamde *Centrale Limiet Stelling*.

1. Het gemiddelde van verschillende steekproefgemiddelen is een goede schatting van het populatiegemiddelde;
2. Hoe groter de omvang van de steekproeven waarop we de verschillende steekproefgemiddelen baseren, des te precieser wordt de schatting op basis van het gemiddelde van de verschillende steekproefgemiddelen.

9.4.4



Opnieuw het voorbeeld van de snoepjesvangst.

- a) In 9.4.2 maakte je een object aan met de naam Gemiddelen1 waarin je alle steekproefgemiddelen op basis van twee kinderen (Steek1-Steek10) wegschrijft.

Maak een histogram aan op basis van Gemiddelen1.

- b) Neem opnieuw tien steekproeven van telkens twee kinderen. Bereken telkens het gemiddelde voor elk van die steekproeven en schrijf deze weg in objecten met de naam Steek21-Steek30. Voeg deze nieuwe objecten samen in een object met Gemiddelen1 en noem dit object Gemiddelen2. Daartoe kan je het volgende commando gebruiken:

```
> Gemiddelen2<-c(Gemiddelen1, Steek21, Steek22, Steek23,
  Steek24, Steek25, Steek26, Steek27, Steek28, Steek29, Steek30)
```

Maak een histogram aan op basis van Gemiddelen2. Vergelijk beide histogrammen. Kan je daar een tendens uit afleiden?

9.4.5



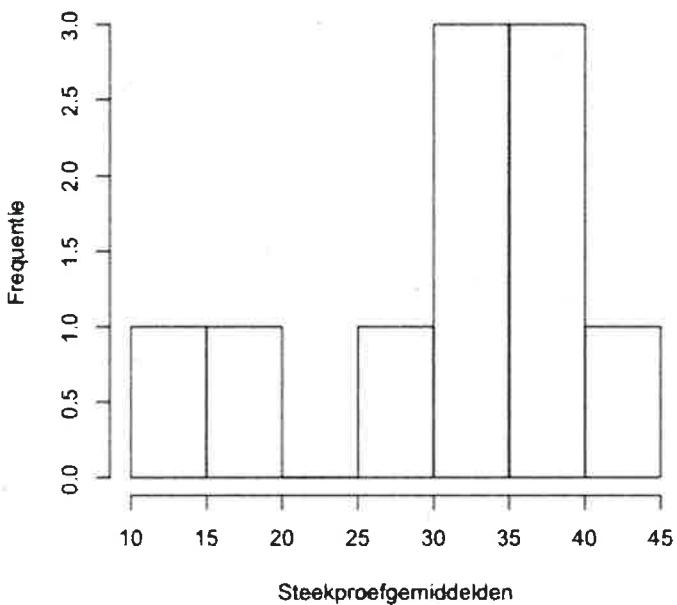
Als je de gemiddelden van alle mogelijke steekproeven uit een populatie onder de vorm van een histogram zou weergeven, dan krijg je een ander soort verdeling: **de verdeling van een parameter – in dit geval het gemid-**

delde – van de populatie. Deze verdeling noemen we de **steekproevenverdeling** van het populatiegemiddelde (*sampling distribution*).

Wat deze steekproevenverdeling van een bepaalde parameter, zoals het gemiddelde of de standaarddeviatie, bijzonder maakt, is dat er kan worden aangetoond dat deze verdeling (al dan niet vertrekende van enkele assumpties over hoe de variabele in de populatie verdeeld is) een of andere theoretische kansverdeling volgt. En naarmate het aantal steekproeven groter is, zal ook de resulterende steekproevenverdeling dichter in de buurt komen van die theoretische kansverdeling.

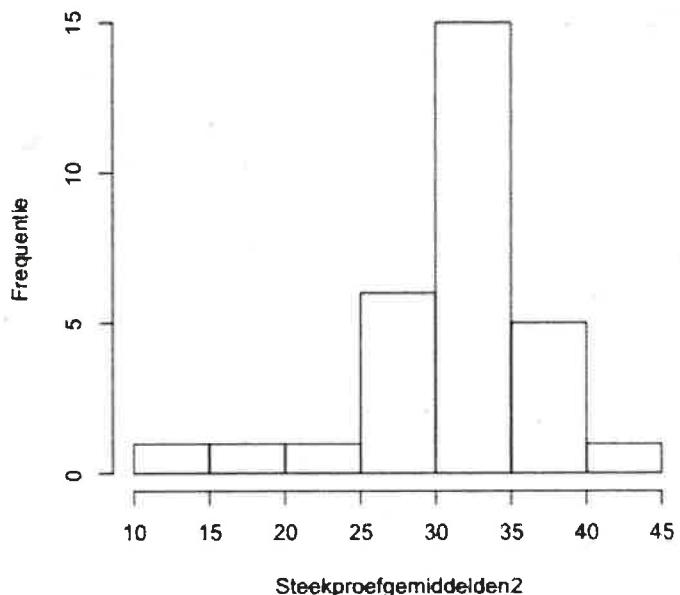
Dit klinkt eerder abstract. Laat het ons toepassen op het gemiddelde als kengetal. In 9.4.4 stelden we vast dat de vorm van de steekproevenverdeling in feite de standaardnormaalverdeling volgt.

Als we de gemiddeldes zouden nemen van 10 steekproeven van de Halloween-kinderen van Zichen-Zussen-Bolder, op basis van 5 kinderen uit de rijen (we selecteren bijvoorbeeld telkens om het andere kind, dus rij 1-kind 1, rij 1-kind 3,...). De verdeling van deze gemiddelden vertoont het volgende verloop:



Figuur 9.1: Steekproevenverdeling op basis van 10 steekproeven ($n=5$)

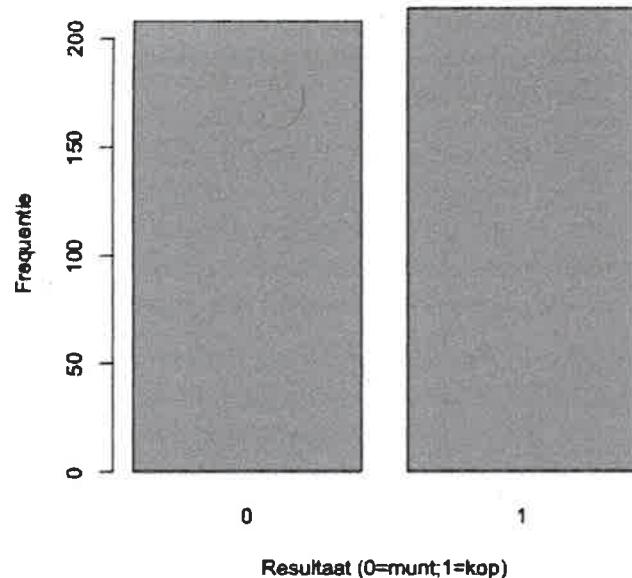
Verhogen we het aantal steekproeven tot bijvoorbeeld 30 (zie figuur 9.2) dan bekomen we een figuur die dichter in de buurt van een normaalverdeling komt en waarbij het gemiddelde dichter bij het populatiegemiddelde komt te liggen.



Figuur 9.2: Steekproevenverdeling op basis van 30 steekproeven ($n=5$)

Deze “wetmatigheid” noemen we de **Centrale Limietstelling**. Voor normaal verdeelde variabelen geldt dat de steekproevenverdeling van het gemiddelde steeds het patroon van de normaalverdeling benadert. Bovendien kan er worden aangetoond dat voor niet-normaal verdeelde variabelen de steekproevenverdeling van het gemiddelde de normaalverdeling meer en meer benadert naarmate de steekproefomvang toeneemt.

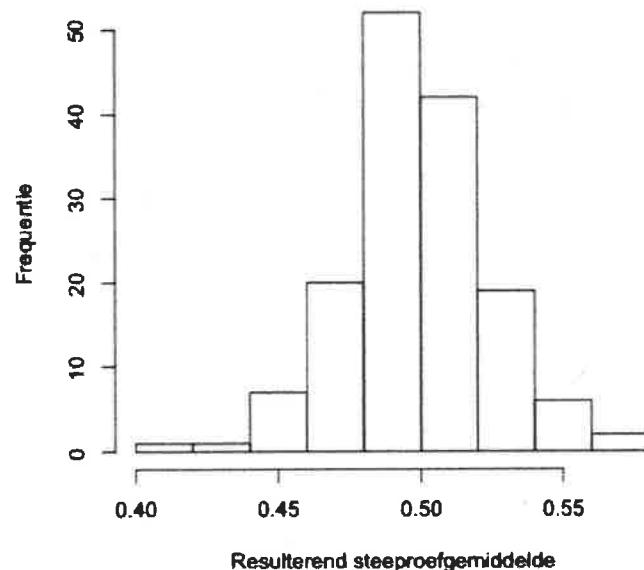
Het opsmijten van een muntje heeft bijvoorbeeld maar twee mogelijke uitkomsten: kruis of munt. Dus als we een 422 keer een muntje opgooien dan krijgen we 422 observaties waarvoor het resultaat bijvoorbeeld 0 is indien munt wordt uitgekomen en 1 indien kop wordt uitgekomen. Bijgevolg is het opgooien van een muntje niet normaal verdeeld. Figuur 9.3 illustreert dit.



Figuur 9.3: Verdeling van het resultaat van 422 keer een muntje opgooien

Stel dat we naar de aanwezigheid van kop kijken (krijgt waarde 1), terwijl de aanwezigheid van munt waarde 0 krijgt, en we herhalen dit opgooien 422 keer. Het gemiddelde van onze steekproef van kop-munt opgooien zal in de buurt van 0,5 komen te liggen (voor de gegevens uit figuur 9.3 bedraagt dit 0,507). Herhaal ik datzelfde experiment (opnieuw 422 keer opgooien) dan zal het gemiddelde steeds licht onder of licht boven 0,5 uitkomen. Het is wel mogelijk dat we een uitzonderlijke keer door toeval verscheidene keren achtereen munt gooien. We krijgen dan een gemiddelde waarde dicht bij 0. Een andere “steekproef” van het opgooien van ons muntje kan ons net zo uitzonderlijk een hele serie van kop achtereen opleveren. Het gemiddeld aantal keren kop zal voor die zeldzame steekproef meer in de buurt van 1 liggen. Maar ook hier gaat die “wetmatigheid” op dat het vaak herhalen van dezelfde steekproef een steekproevenverdeling oplevert die de normaalverdeling benadert. Hieronder vinden we een frequentiehistogram van het gemiddeld aantal keren kop uit 160 steekproeven van telkens 422 keren het opgooien van een muntje (gesimuleerd in R). Deze “steekproevenverdeling” benadert een normaalverdeling. Nemen we het gemiddelde van al deze gemiddelden dan komen we zeer

dicht in de buurt van 0,5 (voor deze 160 steekproeven is het gemiddelde van de gemiddelden 0,499).



Figuur 9.4: Verdeling van de resulterende steekproefgemiddelden van 160 steekproeven van telkens 422 keer een muntje opgooien

Voor andere kengetallen zoals de variantie, standaardafwijking, scheefheid en kurtosis hebben statistici eveneens kunnen aantonen dat de steekproevenverdeling van deze respectievelijke kengetallen onder bepaalde omstandigheden (assumpties) een of andere welbepaalde theoretische kansverdeling volgt. Zo volgt de steekproevenverdeling van de variantie de zogenaamde Chi-kwadraat verdeling indien de variabele in de populatie zelf normaal verdeeld is. Deze eigenschap van de steekproevenverdeling zullen we in de komende hoofdstukken verder hanteren om meer af te kunnen leiden over de waarde van die kengetallen in de populatie.

Samenvattend kunnen we het volgende concluderen:

- 1) We hanteren bijna altijd een steekproef in onderzoek om schattingen te maken over een bepaald fenomeen in de populatie;
- 2) Elke schatting van een kengetal op basis van een steekproef bevat een zekere marge van onjuistheid: de steekproeffout;

- 3) Hoe groter de steekproefomvang is (hoe meer observatie-eenheden), hoe kleiner de steekproeffout (Wet van de Grote Getallen);
- 4) De theoretische verdeling van alle mogelijke schattingen van een kengetal op basis van alle mogelijke steekproeven uit een populatie noemen we de steekproevenverdeling van het welbepaalde kengetal (bv. de steekproevenverdeling van het gemiddelde);
- 5) De steekproevenverdeling van het gemiddelde is altijd bij benadering normaal verdeeld ongeacht de verdeling van de variabele zelf. En hoe groter de steekproefomvang is, hoe dichter de steekproevenverdeling de normaalverdeling benadert. Dit heet de Centrale Limietstelling;
- 6) Voor alle andere kengetallen kan aangetoond worden dat, onder bepaalde assumpties, de steekproevenverdeling van diezelfde kengetallen één of andere theoretische kansverdeling volgt (bv. Chi-kwadraat distributie).

Responsen

Respons 9.1.1

Er zijn verschillende mogelijke personen die je zou kunnen bevragen. Er is daarbij geen juist of fout antwoord. We overlopen een aantal mogelijkheden:

- Ga je dat voorleggen aan de leerlingen van de school die bij jou het dichtst in de buurt ligt?
- Ga je dat aan alle leerlingen secundair onderwijs in Vlaanderen voorleggen? Ook aan de leerlingen in het buitengewoon onderwijs?
- Ga je alle jaren bevragen?
- Ga je alle leerlingen bevragen?
- Misschien kan je je eerder beperken tot de 14- tot 18-jarigen (2de en 3de graad) van het secundair onderwijs? Ga je je beperken tot "gemakkelijke scholen" (zonder probleemgedrag) of net tot "moeilijke scholen" (bv. scholen die in de pers zijn gekomen wegens incidenten)?

Respons 9.1.3

a) Je wilt weten wat de houding is van **leerlingen uit het secundair onderwijs** ten aanzien van geweld op school. In dit geval bestaat de populatie dus uit alle leerlingen uit het secundair onderwijs.

b) Deze populatie is niet zeer precies gedefinieerd. Het kan gaan om leerlingen uit het secundair onderwijs op verschillende momenten (bijvoorbeeld schooljaar 2006-2007 of schooljaar 1997-1998), uit verschillende onderwijsvormen (bijvoorbeeld aso, bso, tso of kso), uit verschillende leerjaren (bijvoorbeeld 1ste jaar uit de 1ste graad of 2de jaar uit de 3de graad),...

Respons 9.2.2

Hoogst waarschijnlijk ga je als volgt te werk:

- je gaat toevallig een aantal balletjes trekken uit het ballenbad
- je gaat turven hoeveel je er hebt van elke kleur
- op basis van wat je geturfd hebt, veronderstel je dat de verhouding in het geheel van het ballenbad vrij identiek is

Een noodzakelijke voorwaarde om die conclusie te trekken is dat je ernaar streeft een representatieve steekproef te trekken die voldoende groot is. Zo zal je bijvoor-

beeld stoppen met het trekken van balletjes uit het ballenbad eens je merkt dat elke extra bal nog weinig verschil uitmaakt in de relatieve aantallen van de verschillende kleuren. En als je nog nieuwe kleuren vindt, ga je waarschijnlijk besluiten dat je nog even verder moet doen.

Respons 9.2.4

Deze steekproef is niet waarheidsgetrouw. In de populatie van Belgen tussen 18 en 75 jaar oud ligt de verhouding tussen mannen en vrouwen beduidend anders. We noemen deze steekproef niet representatief voor de populatie. In dit geval kan dat bovendien verregaande gevolgen hebben voor de bevindingen, want de kans is zeer groot dat mannen heel anders tegen deze vraag aankijken dan vrouwen.

Indien je een willekeurige steekproef trekt is de kans echter zeer klein dat je zo'n extreem niet-representatieve steekproef bekomt.

Respons 9.3.3

a) Aangezien aantal snoepjes een continue maat is, kan je het gemiddeld aantal snoepjes per kind berekenen:

```
> mean(Snoepjes$Snoep, na.rm=TRUE)
[1] 30.96
```

b) Hieronder het gemiddelde voor een steekproef met omvang twee die de volgende aantallen opleverde: 28 en 14. Het resultaat wordt onmiddellijk in Steekproef1 weggeschreven. Vervolgens nemen we van beide getallen het gemiddelde en schrijven daarvan het resultaat weg naar Steek1.

```
> Steekproef1<-sample(Snoepjes$Snoep, 2)
> Steekproef1
[1] 28 14
> Steek1<-mean(Steekproef1)
```

De bovenstaande commando's kunnen eveneens gecombineerd worden tot één commando:

```
> Steek1<-mean(sample(Snoepjes$Snoep, 2))
```

c) Vervolgens herhalen we dit proces negen keer. Echter, we combineren beide commando's in één commando.

```
> Steek2<-mean(sample(Snoepjes$Snoep, 2))
```

```
> Steek3<-mean(sample(Snoepjes$Snoep, 2))
> Steek4<-mean(sample(Snoepjes$Snoep, 2))
> Steek5<-mean(sample(Snoepjes$Snoep, 2))
> Steek6<-mean(sample(Snoepjes$Snoep, 2))
> Steek7<-mean(sample(Snoepjes$Snoep, 2))
> Steek8<-mean(sample(Snoepjes$Snoep, 2))
> Steek9<-mean(sample(Snoepjes$Snoep, 2))
> Steek10<-mean(sample(Snoepjes$Snoep, 2))
```

d) Daarna was het de bedoeling om twee afzonderlijke steekproeven te trekken met telkens een omvang van tien kinderen. Hieronder de bewerkingen om de respectievelijke gemiddeldes uit elk van die twee steekproeven te berekenen en weg te schrijven:

```
> Steek11<-mean(sample(Snoepjes$Snoep, 10))
> Steek12<-mean(sample(Snoepjes$Snoep, 10))
```

e) Tot slot dienden we het gemiddelde van het aantal snoepjes te berekenen op basis van een twee steekproeven van telkens 50 kinderen.

```
> Steek13<-mean(sample(Snoepjes$Snoep, 50))
> Steek14<-mean(sample(Snoepjes$Snoep, 50))
```

f) We kunnen nu alle verschillende gemiddeldes opnieuw oproepen door simpelweg de naam van het object in te typen in R:

```
> Steek1
[1] 21
> Steek2
[1] 38.5
> Steek3
[1] 37.5
> Steek4
[1] 30
> Steek5
[1] 35.5
> Steek6
[1] 45.5
> Steek7
[1] 46.5
> Steek8
[1] 32.5
> Steek9
[1] 21
> Steek10
```

```
[1] 46.5
> Steek11
[1] 31.4
> Steek12
[1] 28.3
> Steek13
[1] 29.6
> Steek14
[1] 30.72
```

Bekijken we alle steekproeven op basis van twee respondenten (Steek1-Steek10) dan merk je grote verschillen in de gemiddelden. Sporadisch ligt er een gemiddelde dicht bij het werkelijke populatiegemiddelde.

De gemiddelden op basis van tien kinderen (Steek11 en Steek 12) zijn beide al dichter bij het werkelijke populatiegemiddelde.

Tot slot merken we dat de twee steekproeven op basis van de helft van de data (Steek13 en Steek14) onderling nog minder verschillen en dichter tegen het werkelijke populatiegemiddelde aanleunen.

Als je alles op een rijtje zet kan je afleiden dat de gemiddelden berekend aan de hand van de grotere steekproeven (die met de meeste waarnemingen) de beste benadering van de ware gemiddelden opleveren en onderling het minst variëren.

Respons 9.4.2

a) Hieronder de commando's voor dit eerste onderdeel:

```
> Gemiddelden1 <- c(Steek1, Steek2, Steek3, Steek4, Steek5, Steek6,+
Steek7, Steek8, Steek9, Steek10)
> Schatting1<-mean(Gemiddelden1)
> Schatting1
[1] 35.45
```

Het gemiddelde van de tien gemiddeldes telkens gebaseerd op een steekproef van twee kinderen bedroeg bij ons 35.45

b) Het gemiddelde van het gemiddelde van 10 steekproeven van 2 individuen benadert het populatiegemiddelde.

```
> Schatting2<-mean(Steek11,Steek12)
> Schatting2
[1] 31.4
```

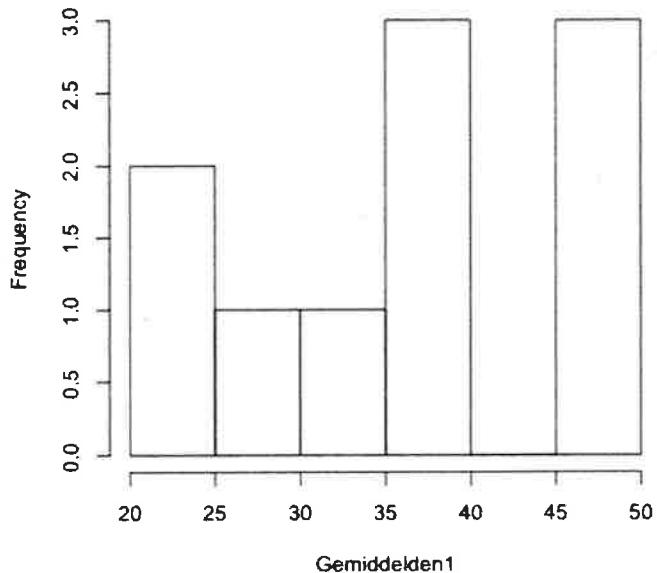
c) Het gemiddelde van de twee gemiddeldes van telkens tien kinderen ligt beduidend dichter bij het werkelijke populatiegemiddelde (30,96) dan het gemiddelde van tien gemiddelden op basis van een steekproef met een omvang van twee kinderen.

Respons 9.4.4

a) Om een histogram aan te maken op basis van de informatie in het object Gemiddelden1 volstaat het volgende commando:

```
> hist(Gemiddelden1)
```

Dit levert het volgende resultaat op:



Figuur 9.5: Verdeling van de tien verschillende gemiddelden gebaseerd op steekproeven van telkens een omvang van twee kinderen

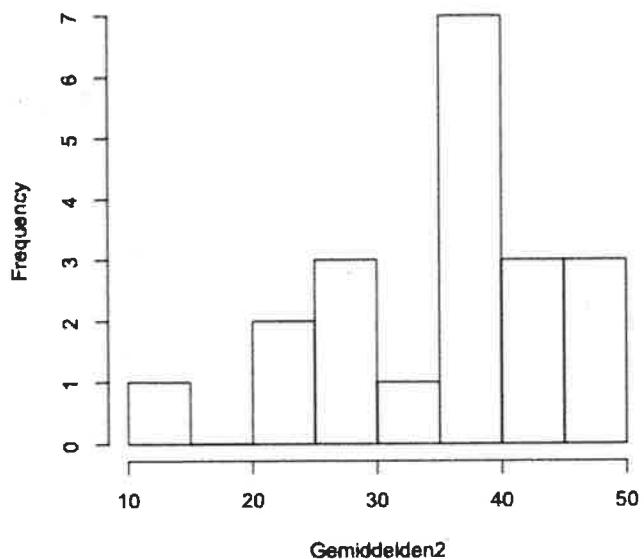
b) Om deze deelopdracht te maken hadden we eerst opnieuw nieuwe steekproeven nodig die telkens uit twee kinderen bestonden. Hieronder de daartoe gehanteerde commando's, met op het einde de commando's om een nieuw object te maken en een histogram:

```
> Steek21<-mean(sample(Snoepjes$Snoep, 2))
> Steek22<-mean(sample(Snoepjes$Snoep, 2))
> Steek23<-mean(sample(Snoepjes$Snoep, 2))
```

```

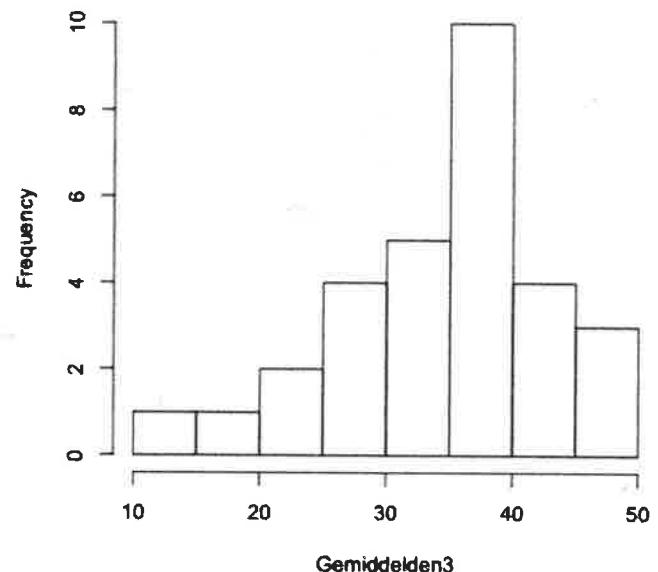
> Steek24<-mean(sample(Snoepjes$Snoep,2))
> Steek25<-mean(sample(Snoepjes$Snoep,2))
> Steek26<-mean(sample(Snoepjes$Snoep,2))
> Steek27<-mean(sample(Snoepjes$Snoep,2))
> Steek28<-mean(sample(Snoepjes$Snoep,2))
> Steek29<-mean(sample(Snoepjes$Snoep,2))
> Steek30<-mean(sample(Snoepjes$Snoep,2))
#
#VERVOLGENS ALLES IN EEN KOLOM ZETTEN SAMEN MET GEGEVENS UIT
#GEMIDDELDEN1
#
> Gemiddelden2<-c(Gemiddelden1, Steek21, Steek22, Steek23, Steek24, +
  Steek25, Steek26, Steek27, Steek28, Steek29, Steek30)
> hist(Gemiddelden2)

```



Figuur 9.6: Verdeling van de 20 verschillende gemiddelden gebaseerd op steekproeven van telkens een omvang van twee kinderen

Indien we beide histogrammen bekijken dan merken we dat het tweede histogram al iets meer een piek gaat vertonen dan het eerste histogram. Met een beetje goede wil, kan je stellen dat het tweede histogram er al iets meer gaat uitzien als een normaalverdeling. Indien we nog eens tien steekproeven extra toevoegen dan begint deze tendens zich nog duidelijker af te tekenen. Hieronder ter illustratie een histogram gebaseerd op 30 steekproeven van telkens een omvang van twee kinderen.



Figuur 9.7: Verdeling van de 30 verschillende gemiddelden gebaseerd op steekproeven van telkens een omvang van twee kinderen

Gehanteerde functies

Functie	Doelstelling	Bron
<code>sample(x, n)</code>	Neemt een steekproef van waarden uit alle waarden voor een kolom in R. Tussen de haakjes dien je enerzijds te verwijzen naar de kolom waaruit een steekproef van waarden moet worden genomen (x) en anderzijds de steekproefomvang meegeven (n).	R basispakket

HOOFDSTUK 10

Inferenties over de verdeling van variabelen in de populatie

DOELSTELLINGEN:

Na dit hoofdstuk

- weet je hoe de standaardfout van het gemiddelde berekend wordt;
- weet je hoe je op basis van de standaardfout van het gemiddelde een betrouwbaarheidsinterval kan berekenen;
- kan je een betrouwbaarheidsinterval voor het gemiddelde in R berekenen;
- weet je wat de Chi-kwadraatverdeling is;
- weet je hoe je op basis van de Chi-kwadraatverdeling een betrouwbaarheidsinterval berekent voor de variantie;
- kan je een betrouwbaarheidsinterval voor de variantie in R berekenen;
- weet je hoe je de standaardfout van zowel scheefheid als kurtosis kan berekenen;
- kan je de standaardfout van zowel scheefheid als kurtosis in R berekenen;
- weet je hoe je op basis van de standaardfout van zowel scheefheid als kurtosis een betrouwbaarheidsinterval kan berekenen;
- kan je een betrouwbaarheidsinterval voor zowel scheefheid als kurtosis in R berekenen.

NODIGE FILES:

Pirls2.RData

OLP Functies.R

een file met daarin aangepaste functies die bij dit OLP horen



In de vorige hoofdstukken introduceerden we beschrijvende kengetallen om een beeld te vormen van de ligging, spreiding en vorm van de verdeling van een variabele. Deze kengetallen hebben we, zonder daarbij enige kanttekeningen te plaatsen, toegepast op steekproefgegevens. In hoofdstuk 9 stelden we echter dat elke steekproef op één of andere wijze leidt tot een niet perfecte schatting van kengetallen. We introduceerden de basisingrediënten van de steekproeftheorie door de concepten "steekproevenverdeling" en "centrale limietstelling" te behandelen. In dit hoofdstuk passen we de steekproeftheorie toe op de verschillende kengetallen die we eerder behandelden. Belangrijk daarbij is dat we dit beperken tot kengetallen die opgaan voor kwantitatieve variabelen.

10.1. Betrouwbaarheidsintervallen rond het gemiddelde

10.1.1

Op basis van de steekproeftheorie weten we dat een steekproefgemiddelde een benadering / een schatting is van een populatiegemiddelde. Door het gemiddelde te berekenen op basis van steekproefgegevens kunnen we een schatting maken aangaande het gemiddelde in een populatie. Enkel, we weten nog niet hoe goed onze schatting is. In welke mate is het steekproefgemiddelde een goede schatting van het populatiegemiddelde? De toepassing van de steekproeftheorie levert ons de nodige bagage om deze vraag beter te kunnen beantwoorden.

Aangezien het steekproefgemiddelde één welbepaalde waarde is, kan je deze ook beschouwen als één van alle mogelijke steekproefgemiddeldes. Het is één van de mogelijke uitkomsten van de variabele waarvan we de verdeling steekproevenverdeling hebben genoemd. Een steekproefgemiddelde wordt ook wel eens omschreven als een **puntschatting** van een populatiegemiddelde. Het is één welbepaalde waarde van een nieuwe variabele, en daardoor weten we eigenlijk niet of het een over- dan wel een onderschatting is van het werkelijke populatiegemiddelde.

Het zou eigenlijk heel handig zijn om het populatiegemiddelde te kennen. Voor informatie zoals de gemiddelde leeftijd van de Belgische bevolking beschikken we over populatiegegevens die met veel moeite zijn verzameld. Maar voor variabelen zoals IQ van alle 18-jarige Vlaamse kinderen hebben we geen populatiegegevens en hebben we bijgevolg het gissen naar het werkelijke populatiegemiddelde.

Daarom dat we beroep doen op een **betrouwbaarheidsinterval** wat ons een indicatie geeft van de grenzen waartussen het werkelijke populatiegemiddelde met een bepaalde betrouwbaarheid ligt.

Betrouwbaarheidsintervallen gaan ons een schatting geven waar het populatiegemiddelde kan liggen onder de vorm van een reikwijdte van waarden. Een betrouwbaarheidsinterval geeft met andere woorden een serie scores waarbinnen we veel kans hebben om dat onbekende reële populatiegemiddelde te vinden. Maar hoe kunnen we zo'n intervallen gaan bepalen? Waarop baseren we ons dan?

10.1.2

 Een vertrekpunt om op deze vraag te antwoorden is de Centrale Limietstelling. Herneem uit hoofdstuk 9 de essentie van de Centrale Limietstelling. Wat is de kernboodschap van die stelling uit de theoretische statistiek?

10.1.3

 Het is het eerste deel van de Centrale Limietstelling dat ons kan helpen om betrouwbaarheidsintervallen te gaan “bepalen”: de steekproevenverdeling van het gemiddelde benadert de normaalverdeling.

Als we dus de steekproevenverdeling zouden kennen, dan kunnen we gebruik maken van de eigenschappen van de normaalverdeling: de 68-95-99,7 regel. Echter, de steekproevenverdeling zullen we nooit “kennen”. Hiervoor moeten we heel veel steekproeven nemen uit de populatie, wat we nooit in de realiteit doen. We kunnen wel de steekproevenverdeling “inschatten”.

10.1.4

 Door gebruik te maken van de Centrale Limietstelling kunnen we een benadering maken van de steekproevenverdeling van het gemiddelde. Daartoe hebben we bepaalde kengetallen nodig die de eigenschappen van deze steekproevenverdeling samenvatten. Welke kengetallen hebben we nodig om de steekproevenverdeling te kennen?

10.1.5

 Wiskundige statistici hebben voor ons onderzoek gedaan naar wat de beste schattingen zijn van zowel het gemiddelde als de standaardafwijking van de steekproevenverdeling van het gemiddelde. We spreken hier van schattingen aangezien we als onderzoeker enkel gebruik kunnen maken

van wat onze ene steekproef ons leert. Hieronder ontrafelen we ter illustratie de redenering (theorie) die deze statistici hebben opgebouwd om beide “onbekenden” in te schatten. Om heel die redenering uit de doeken te doen, dienen we dezelfde taal te hanteren. Daarom maken we gebruik van enkele conventies/notaties:

- De verdeling van een variabele in een populatie wordt met een hoofdletter geschreven. X verwijst bijvoorbeeld naar de verdeling van variabele X in de hele populatie.
- Ook kengetallen in de populatie worden met een hoofdletter geschreven. Bijvoorbeeld het gemiddelde van X noteren we als \bar{X} .
- Kengetallen in de steekproef worden met een kleine letter geschreven. Zo noteren we het gemiddelde van x als \bar{x} .
- Indien we willen verwijzen naar de steekproevenverdeling van een bepaald kengetal dan hanteren we de notatie van dat kengetal. Stel dat we bijvoorbeeld willen verwijzen naar de variantie van de steekproevenverdeling van het gemiddelde dan noteren we dit als $VAR(\bar{X})$.
- In statistische taal worden schattingen altijd neergeschreven als $E()$. Willen we bijvoorbeeld verwijzen naar de schatting van het populatiegemiddelde dan schrijven we $E(\bar{X})$.

Een eerste bouwsteen van de theorie is dat een steekproefgemiddelde (\bar{x}) de beste schatting is van het populatiegemiddelde (\bar{X}). Bovendien is het gemiddelde van de steekproevenverdeling (\bar{x}) gelijk aan het populatiegemiddelde (\bar{X}). Formeel geeft dit het volgende:

$$E(\bar{X}) = \bar{x} = \bar{\bar{X}}$$

Een eerste “onbekende” om de steekproevenverdeling in te schatten was het gemiddelde van de steekproevenverdeling (\bar{x}). Op basis van de bovenstaande formule weten we dat het steekproefgemiddelde op zich de beste schatting is van de steekproevenverdeling.

Stel dat we van 450 kinderen een IQ-toets hebben afgenomen en gemiddeld scoren deze leerlingen 100 op de IQ-toets. Dan is 100 ook de beste schatting van het gemiddelde van alle mogelijke steekproeven uit de steekproevenverdeling van het gemiddelde en bijgevolg ook de beste schatting van het populatiegemiddelde.

De tweede bouwsteen van de theorie is dat de variantie van de steekproevenverdeling van het gemiddelde gelijk is aan de variantie in de populatie

gedeeld door de steekproefomvang. Dit kunnen we in de onderstaande formule gieten

$$VAR(\bar{x}) = \frac{VAR(X)}{n}$$

waarbij n staat voor de steekproefomvang. Echter, veel verder zijn we hiermee nog niet want de variantie van onze variabele X in de hele populatie is op zich ook een "onbekend" gegeven. Statistici hebben echter aangetoond dat deze variantie benaderd kan worden door de variantie in de steekproef (s_x^2) vermenigvuldigd met de ratio $(N-1)/N$ waarbij N staat voor de populatieomvang. In een formule geeft dit:

$$VAR(X) = \left(\frac{N-1}{N}\right)(s_x^2)$$

- 10.1.6** Hernemen we het voorbeeld van de IQ-scores. Stel dat we weten dat de populatie bestaat uit 101825 kinderen van vergelijkbare leeftijd en dat in de steekproef de variantie 225 bedraagt. Wat is dan de beste schatting van de variantie in de populatie?

- 10.1.7** Het hoeft geen uitvoerig betoog om aan te tonen dat hoe groter de populatie wordt, hoe dichter de factor $((N-1)/N)$ de waarde één benadert. Daarom wordt door statistici vaak aangenomen dat de variantie die we vaststellen in de steekproef een vrij goede schatting is van de variantie in de populatie.

Herneem de formule van de variantie in de steekproevenverdeling. Hoe kunnen we nu de "onbekende waarde" in die formule inschatten? Herschrijf de formule met die nieuwe schatting.

Pas de formule die je uitkomt toe op het voorbeeld met IQ-scores.

- 10.1.8** De variantie is echter niet het meest geschikte kengetal van spreiding aangezien deze de spreiding niet uitdrukt in de originele schaal van de variabele. Het gaat om een uitdrukking van spreiding in gekwadrateerde vorm. Daarom introduceerden we eerder het kengetal "de standaardafwijking". Dit is niet meer of niet minder dan de vierkantswortel van de variantie. Dus, willen we de standaardafwijking van de steekproevenverdeling ken-

nen, dan nemen we de vierkantswortel van de variantie in de steekproevenverdeling. De onderstaande formule geeft dit weer:

$$SD(\bar{x}) = \sqrt{\frac{s_x^2}{n}} = \frac{s_x}{\sqrt{n}}$$

De standaardafwijking van de steekproevenverdeling is met andere woorden gelijk aan de standaardafwijking in de steekproef gedeeld door de vierkantswortel van de steekproefomvang. Deze grootheid, de standaardafwijking van de steekproevenverdeling, wordt doorgaans bestempeld als **de Standaardfout voor het gemiddelde (Standard Error of the mean)** in de statistiek.

- 10.1.9** Herneem het voorbeeld van de IQ-scores. Bereken de standaardfout voor het gemiddelde.

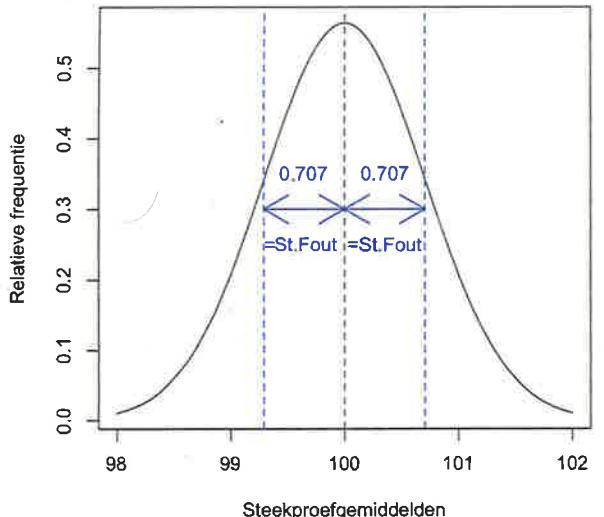


- 10.1.10** We vatten even samen. We zijn op zoek naar hoe de steekproevenverdeling voor het kengetal het gemiddelde eruit ziet. Nu weten we dat deze de normaalverdeling volgt met als gemiddelde waarde het steekproefgemiddelde en met als variantie de steekproefvariantie gedeeld door de steekproefomvang.

We hernemen het voorbeeld van IQ-scores. Het ging daarbij om een steekproef van 450 leerlingen die een gemiddeld IQ hadden van 100 en een variantie van 225 vertoonden. Op basis van die steekproefgegevens kunnen we ook de steekproevenverdeling inschatten:

- deze volgt de normaalverdeling;
- met als gemiddelde 100;
- met als standaardafwijking $0,707 (= \sqrt{225}/\sqrt{450})$, ook wel standaardfout genoemd.

Figuur 10.1 geeft deze steekproevenverdeling grafisch weer.



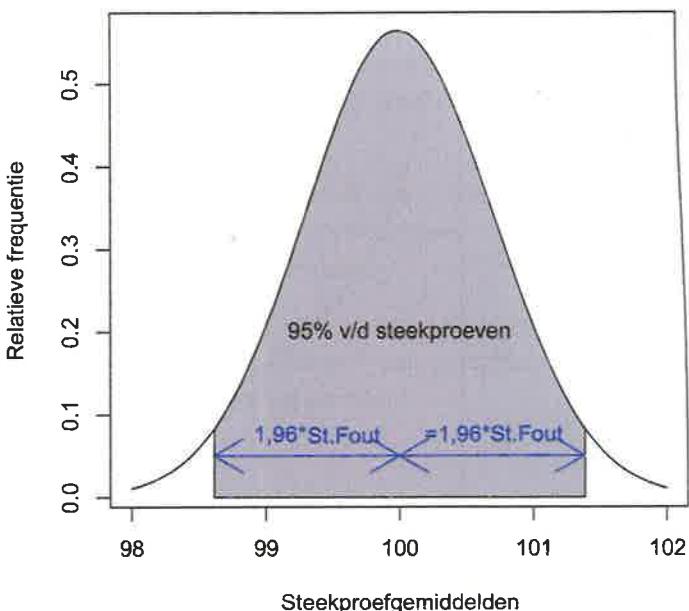
Figuur 10.1: Steekproevenverdeling van het gemiddelde voor het IQ voorbeeld

10.1.11 Aan het begin van dit betoog stelden we dat het gemiddelde van een steekproef een “puntschatter” was van het populatiegemiddelde. Hoe kunnen we gebruik maken van de steekproevenverdeling van het gemiddelde om het populatiegemiddelde beter te kunnen inschatten?

10.1.12 Op basis van de steekproevenverdeling kunnen we de zogenaamde **betrouwbaarheidsintervallen** bepalen. Door gebruik te maken van de eigenschappen van de normaalverdeling kunnen we een onder- en boven-grens bepalen waartussen het ware populatiegemiddelde met 95% betrouwbaarheid ligt. De onderstaande figuur geeft dit weer voor het IQ- voorbeeld. Door het steekproefgemiddelde te verminderen met 1,96 keer (afgerond 2 keer, zie respons 10.3.2) de standaardfout verkrijgen we de ondergrens en door het steekproefgemiddelde te vermeerderen met 1,96 keer de standaardfout krijgen we de boven-grens van het 95% betrouwbaarheidsinterval. Het 95% betrouwbaarheidsinterval voor het IQ- voorbeeld heeft als ondergrens $98,614 (= 100 - (1,96 * 0,707))$ en als boven-grens $101,386 (= 100 + (1,96 * 0,707))$.

Het 95% betrouwbaarheidsinterval voor het populatiegemiddelde wordt dus bepaald door de formule:

$$\bar{x} \pm 1,96 * \frac{s}{\sqrt{n}}$$



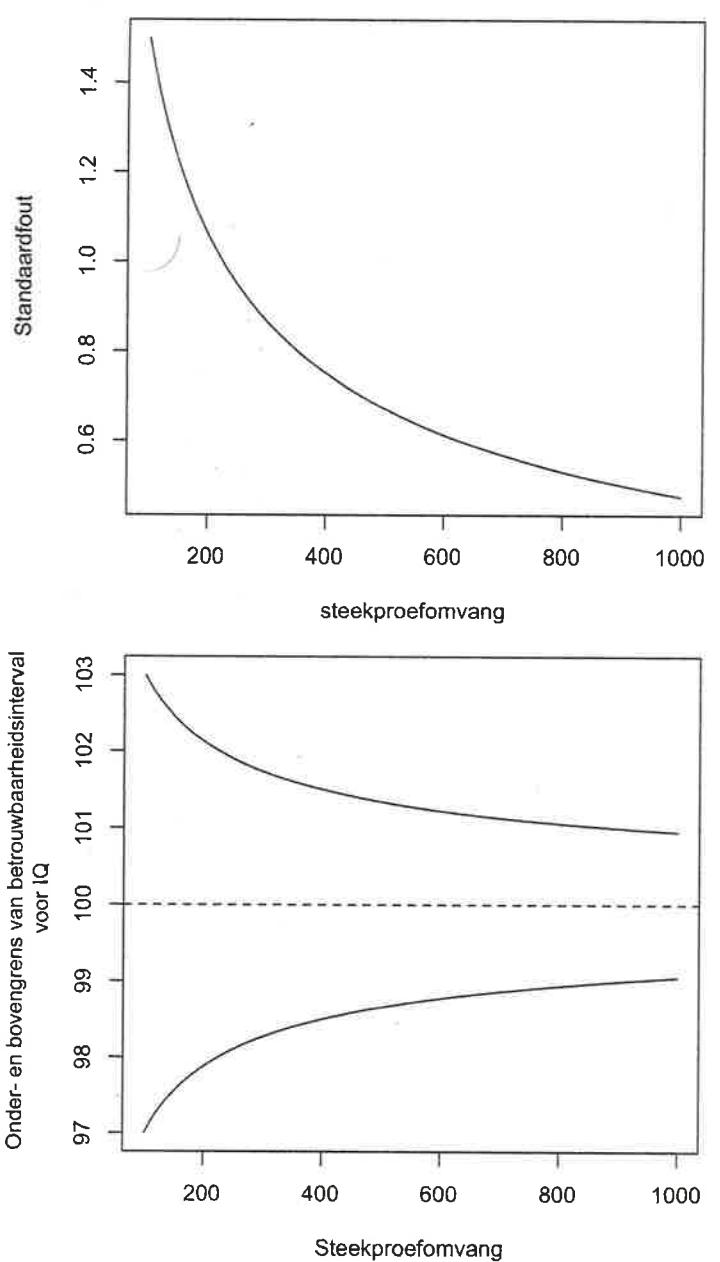
Figuur 10.2: Illustratie van 95% betrouwbaarheidsinterval voor IQ-voorb

10.1.13 We hernemen het voorbeeld van de IQ-scores. Bereken voor elk van de volgende situaties de betrouwbaarheidsintervallen:

- a) steekproefgemiddelde=100, standaardafwijking=15, n=225;
- b) steekproefgemiddelde=100, standaardafwijking=15, n=450;
- c) steekproefgemiddelde=100, standaardafwijking=15, n=675.

Wat kan je afleiden uit deze drie berekeningen?

10.1.14 Uit de bovenstaande oefening kunnen we afleiden dat hoe groter de steekproef wordt hoe smaller de 95% betrouwbaarheidsintervallen worden. Grote steekproeven leiden met andere woorden tot preciezere schattingen. Dit wordt in figuur 10.3 geïllustreerd aan de hand van het IQ- voorbeeld. Naarmate de steekproefomvang groter wordt, daalt de steekproeffout (bovenste figuur) en worden het betrouwbaarheidsinterval smaller (onderste figuur).



Figuur 1.3: De relatie tussen steekproefomvang en steekproeffout en de relatie tussen steekproefomvang en betrouwbaarheidsinterval toegepast op het IQ-voorbeeld

10.1.15 Tot hertoe hebben we de theorie achter betrouwbaarheidsintervallen besproken en pasten we dit toe zonder met werkelijke data te werken. In R kunnen we die theorie toepassen. Om het rekenwerk te vergemakkelijken voegden we in het pakket “OLP functies.R” enkele nuttige functies toe.

De eerste functie heet `standaardfout()` en berekent de standaardfout van het gemiddelde uit voor een variabele. Je hoeft tussen de haakjes enkel te verwijzen naar een variabele.

Daarnaast hebben we de functie `betr.interval()` opgenomen. Door tussen haakjes te verwijzen naar een variabele wordt standaard het 95%-betrouwbaarheidsinterval berekend.

> `betr.interval(x)`

Deze functie kunnen we ook gebruiken om een ander betrouwbaarheidsinterval te berekenen. Daartoe dienen we het extra argument `conf.level = hanteren`. Zo geeft de onderstaande code het 80% berouwbaarheidsinterval van de variabele X:

> `betr.interval(X, conf.level=0.80)`

10.1.16 Het databestand Pirls2.Rdata zijn afkomstig van de internationale PIRLS studie (2006): een vergelijkend onderzoek naar leesprestaties en lesonderwijs in het basisonderwijs. In dit databestand zijn een selectie van variabelen uit het leerlingenbestand opgenomen. Zo bevat dit bestand o.a. de volgende twee variabelen die scores bevatten op twee dimensies van leesvaardigheid:

- Informsscore (leesvaardigheid m.b.t. informatieve teksten);
- Literatuurscore (leesvaardigheid m.b.t. literaire teksten).

Beide variabelen zijn op een gelijkaardige schaal uitgedrukt, wat maakt dat scores met elkaar vergeleken kunnen worden.

- a) Bereken een 95%-betrouwbaarheidsinterval voor beide variabelen door zelf alle stappen van de berekening uit te voeren;
- b) Controleer je uitkomst uit a) aan de hand van de functie `betr.interval()`;
- c) Wat kan je concluderen uit de vergelijking van beide intervallen?

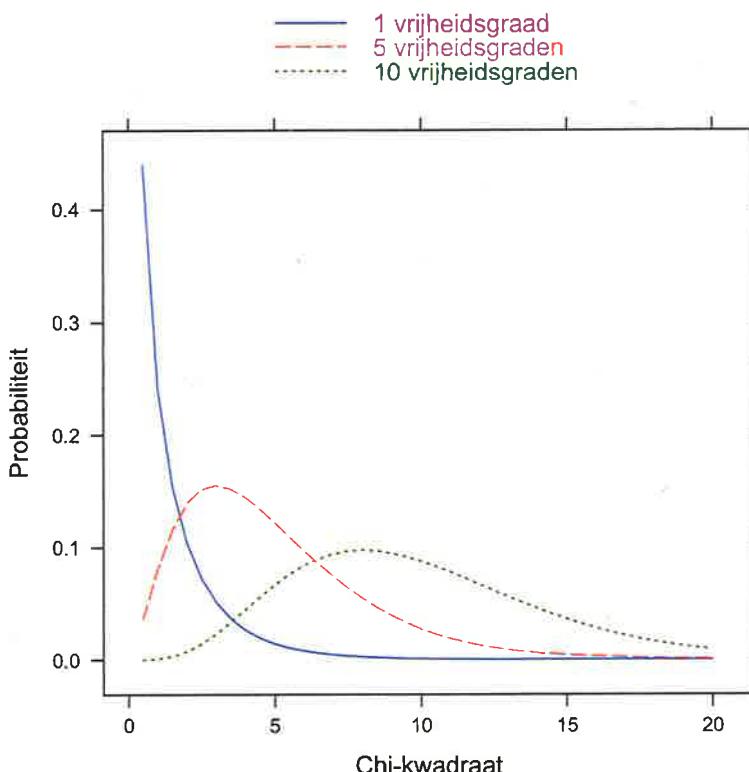
10.2. Betrouwbaarheidsintervallen rond de variantie

- 10.2.1** We bespraken hierboven vrij uitgebreid de redenering achter betrouwbaarheidsintervallen voor het gemiddelde. Een gelijkaardige redenering gaat ook op voor betrouwbaarheidsintervallen rond de variantie.



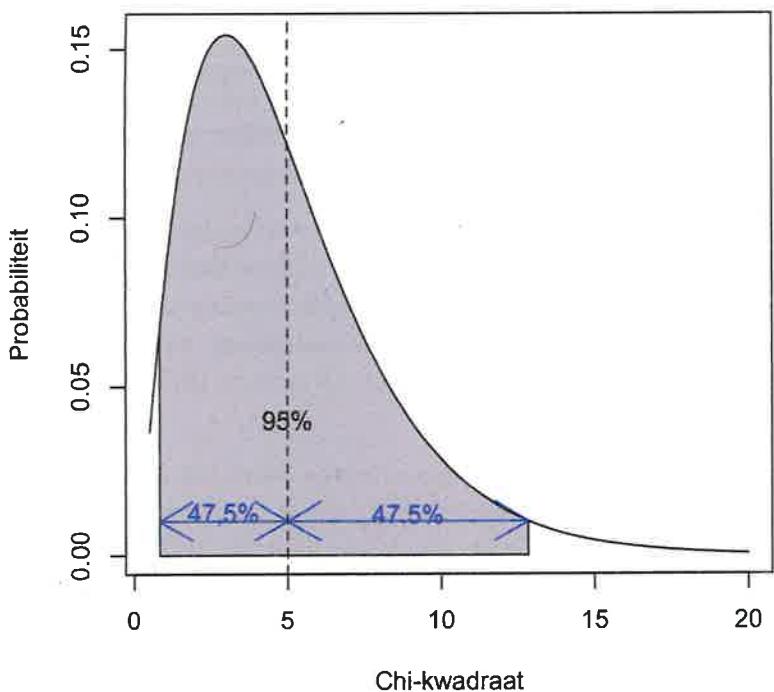
Op basis van een steekproef kunnen we een schatting maken van de variantie voor een bepaald kenmerk in de populatie. Het gaat ook hier om een puntschatter. Elke andere steekproef zou een andere schatting van de variantie opleveren. De verdeling van al deze mogelijke andere schattingen van de variantie noemen we dan de steekproevenverdeling van de variantie.

Voor het gemiddelde konden we ons baseren op de Centrale Limietstelling om te concluderen dat de steekproevenverdeling de standaardnormaalverdeling volgt. Bij de variantie gaat dit niet op. Echter, de steekproevenverdeling van de variantie volgt een andere theoretische kansverdeling: de Chi-kwadraat verdeling. De Chi-kwadraat verdeling is eigenlijk een familie van theoretische kansverdelingen. We kunnen dus verschillende Chi-kwadraatverdelingen onderscheiden die onderling enkel verschillen op basis van wat we "vrijheidsgraden" (degrees of freedom) gaan noemen. In figuur 10.4 hebben we exemplarisch drie verschillende Chi-kwadraatverdelingen voorgesteld: één met slechts 1 vrijheidsgraad; één met 5 vrijheidsgraden; en één met 10 vrijheidsgraden. Kenmerkend voor de Chi-kwadraatverdeling is dat dit geen symmetrische maar een rechtscheve verdeling is. Het gemiddelde van elke Chi-kwadraatverdeling is altijd gelijk aan het aantal vrijheidsgraden en de variantie van de Chi-kwadraatverdeling is gelijk aan twee keer het aantal vrijheidsgraden.



Figuur 10.4: Drie verschillende Chi-kwadraatverdelingen

Analoog aan z-scores kan je bij een Chi-kwadraat verdeling met een bepaald aantal vrijheidsgraden de kans berekenen dat je een bepaalde Chi-kwadraatwaarde terugvindt. Hierover zijn er opnieuw een heleboel tabellen terug te vinden op verschillende websites. Bijgevolg kan je ook opzoeken bij welke Chi-kwadraatwaarde de kans 2,5% bedraagt om die welbepaalde waarde of lager uit te komen. De onderstaande figuur illustreert dit voor de Chi-kwadraatverdeling met 5 vrijheidsgraden. 95% van de waarden ligt tussen 0,83 en 12,83.



Figuur 10.5: Chi-kwadraatverdeling met 5 vrijheidsgraden en 95% interval

We kunnen van deze familie van Chi-kwadraatverdelingen gebruik maken om een 95% betrouwbaarheidsinterval op te maken voor de variantie. Echter, dit gaat niet altijd. We kunnen van deze Chi-kwadraatverdelingen enkel gebruik maken als onze variabele waarvoor we dit willen doen bij benadering normaal verdeeld is. Dus indien we een variabele IQ hebben met een zekere geschatte variantie in onze steekproef (bv. 225) en deze variabele is bij benadering normaal verdeeld, dan kunnen we een betrouwbaarheidsinterval opstellen waarover we met een zekere mate van betrouwbaarheid kunnen zeggen dat de populatievariantie erin zal liggen.

Het bepalen van dit betrouwbaarheidsinterval is minder eenvoudig dan bij het gemiddelde. Zonder verder in te gaan op de theorie hierachter kunnen we meegeven dat we de **ondergrens van een 95% betrouwbaarheidsinterval** als volgt kunnen bepalen:

$$\frac{(n-1)s^2}{\chi_{(n-1; 0,975)}^2}$$

Daarbij staat n voor de steekproefomvang, s^2 voor de variantie in de steekproef en $\chi_{(n-1; 0,975)}^2$ voor de waarde in de Chi-kwadraat verdeling met $n - 1$ vrijheidsgraden die overeenstemt met een cumulatieve kans van 97,5%.

De **bovengrens van een 95% betrouwbaarheidsinterval** kunnen we als volgt bepalen:

$$\frac{(n-1)s^2}{\chi_{(n-1; 0,025)}^2}$$

Met n voor de steekproefomvang, s^2 voor de variantie in de steekproef en $\chi_{(n-1; 0,025)}^2$ voor de waarde in de Chi-kwadraat verdeling met $n - 1$ vrijheidsgraden die overeenstemt met een cumulatieve kans van 2,5%.

Beide formules kunnen we ook generieker toepassen om betrouwbaarheidsintervallen met een andere breedte te berekenen. Zo kan je bijvoorbeeld het 80% betrouwbaarheidsinterval bepalen door enkel in de noemer een aangepaste Chi-kwadraatwaarde te gebruiken: voor de ondergrens de Chi-kwadraatwaarde die overeenstemt met een cumulatieve probabilitet van 90% en voor de bovengrens de Chi-kwadraatwaarde die overeenstemt met een cumulatieve probabilitet van 10%.

10.2.2

 Zowel om de bovengrens als om de ondergrens te bepalen, moeten we op zoek naar de waarde die overeenstemt met een cumulatieve kans van respectievelijk 97,5% en 2,5% binnen een bepaalde specifieke Chi-kwadraatverdeling. Daartoe kunnen we gebruik maken van standaardtabellen. Echter, in R kunnen we dit veel gemakkelijker opvragen aan de hand van de functie `qchisq(p, df)`. Als argumenten voor deze functie geef je enerzijds de kans (p ; een getal tussen nul en 1) en anderzijds het aantal vrijheidsgraden (df) op. Zo krijg je aan de hand van het onderstaande voorbeeld de Chi-kwadraatwaarde die overeenstemt met een cumulatieve probabilitet van 2,5% in de Chi-kwadraatverdeling met 5 vrijheidsgraden:

```
> qchisq(0.025, 5)
[1] 0.8312116
```

In formulevorm geeft dit $\chi_{(5; 0,025)}^2 = 0,8312$

10.2.3 Eerder dit hoofdstuk hanteerden we het voorbeeld van de IQ-scores van 450 leerlingen, met een gemiddelde van 100 en een variantie van 225.



a) Bepaal het 95%-betrouwbaarheidsinterval voor de variantie op basis van de bovenstaande formules en de `qchisq()` functie in R.

b) Bepaal het 80%-betrouwbaarheidsinterval voor de variantie op basis van de bovenstaande formules en de `qchisq()` functie in R.

10.2.4 Aangezien het rekenwerk zoals beschreven in 10.2.3 een aantal tussen-



stappen bevat, vergt het redelijk veel werk om betrouwbaarheidsintervallen te berekenen voor de variantie. Daarom hebben we in "OLP functies.R" twee functies opgenomen die het rekenwerk voor ons doen: `betr.interval.Var1()` en `betr.interval.Var2()`. Het verschil tussen beide functies is dat je bij de eerste functie kan verwijzen naar variabelen in een dataset, bij de tweede geef je zelf aan wat de variantie en het aantal steekproevenheden is. De tweede functie kan je met andere woorden hanteren indien je een betrouwbaarheidsinterval wil reconstrueren voor de variantie, zonder dat je over de eigenlijke data beschikt. De eerste functie hanteer je indien je op eigen data werkt.

De eerste functie bevat volgende elementen die je kan opgeven: de betrokken variabele en eventueel het betrouwbaarheidsniveau (een getal tussen nul en één wat verwijst naar de proportie tussen beide grenzen). Als je zelf geen betrouwbaarheidsniveau opgeeft, wordt standaard het 95% betrouwbaarheidsinterval berekend. Hieronder een voorbeeld van beide.

```
# Berekenen van een 95% betrouwbaarheidsinterval voor een
# fictieve variabele V1 uit een fictieve dataset Data1.
> betr.interval.Var1(Data1$V1)
```

```
# Berekenen van een 80% betrouwbaarheidsinterval voor een
# fictieve variabele V1 uit een fictieve dataset Data1.
> betr.interval.Var1(Data1$V1, conf.level=0.80)
```

De tweede functie vraagt meerdere argumenten: de variantie, de steekproefomvang en eventueel het betrouwbaarheidsniveau. Net als bij de eerder variant wordt er standaard uitgegaan van een 95% betrouwbaarheidsniveau.

```
# Berekenen van een 95% betrouwbaarheidsinterval voor een
# variantie die 200 bedraagt in een steekproef van 1000
# waarnemingen.
> betr.interval.Var2(200,1000)
```

```
# Berekenen van een 80% betrouwbaarheidsinterval voor een
# variantie die 200 bedraagt in een steekproef van 1000
# waarnemingen.
> betr.interval.Var2(200,1000, conf.level=0.80)
```

10.2.5



In 10.2.3 berekende je zelf betrouwbaarheidsintervallen voor het IQ-voorb

beeld.

- a) Bepaal opnieuw het 95%-betrouwbaarheidsinterval voor de variantie.
- b) Bepaal opnieuw het 80%-betrouwbaarheidsinterval voor de variantie.

10.2.6



We hernemen het databestand Pirls2.Rdata en meer specifiek de variabelen **Informscore** (leesvaardigheid m.b.t. informatieve teksten) en **Literatuurscore** (leesvaardigheid m.b.t. literaire teksten).

Herinner je dat beide variabelen op een gelijkaardige schaal uitgedrukt zijn, wat maakt dat scores met elkaar vergeleken kunnen worden.

- a) Bereken de varianties voor beide variabelen;
- b) Bereken een 95%-betrouwbaarheidsinterval voor de variantie van beide variabelen door zelf de verschillende R-commando's toe te passen;
- c) Wat kan je concluderen uit de vergelijking van beide intervallen?

10.2.7



In heel deze uiteenzetting hebben we ons toegespitst op de variantie als kengetal. Eerder in dit OLP stelden we echter dat de variantie op zich moeilijk te interpreteren is aangezien deze de spreiding uitdrukt in de meetschaal van de variabele in het kwadraat. We introduceerden in dit OLP daarom de standaardafwijking, welke gelijk is aan de vierkantwortel van de variantie. Nu, we kunnen betrouwbaarheidsintervallen voor de variantie opnieuw gaan herberekenen tot betrouwbaarheidsintervallen voor de standaardafwijking. Dit doen we door van zowel de onder- als de bovengrens de vierkantwortel te nemen.

10.3. Betrouwbaarheidsintervallen voor de kengetallen van vorm

- 10.3.1** Naast kengetallen voor ligging en spreiding bespraken we ook kengetallen van vorm. Meerbepaald hadden we het over de scheefheid (skewness) en de kurtosis.

Voor beide kengetallen gaat dezelfde redenering op als bij de eerder besproken kengetallen. Als we deze kengetallen berekenen op basis van een steekproef, dan is het resultaat een schatting van de scheefheid en kurtosis van de verdeling in de populatie. Dit roept de vraag op of het gaat om een goede schatting. Betrouwbaarheidsintervallen geven opnieuw meer informatie over de waarde van deze kengetallen in de populatie.

Statistici hebben aangetoond dat de steekproevenverdeling van zowel de scheefheid als de kurtosis de normaalverdeling volgt (net zoals dat het geval was bij het gemiddelde). Indien we weten wat de standaardafwijkingen zijn van deze steekproevenverdelingen, dan hebben we opnieuw voldoende informatie om betrouwbaarheidsintervallen op te stellen. Immers, we kunnen dan opnieuw beroep doen op de eigenschappen van de standaardnormaalverdeling. De standaardafwijkingen van de steekproevenverdelingen voor beide kengetallen gaven we eerder al de naam "standaardfout". Hieronder geven we voor de volledigheid de formules van zowel de standaardfout van de scheefheid (SFS afgekort) als de standaardfout van de kurtosis (SFK afgekort). Beide formules hebben als enige onbekende de steekproefomvang (n).

$$SFS = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

$$SFK = 2 * SFS * \sqrt{\frac{n^2 - 1}{(n-3)(n+5)}}$$

Laat ons beide formules illustreren aan de hand van het voorbeeld van de IQ-scores dat we eerder al hanteerden. Het ging daarbij om een steekproef van 450 kinderen. De berekening van beide standaardfouten ziet er dan als volgt uit:

$$SFS = \sqrt{\frac{6 * 450 * (450-1)}{(450-2)(450+1)(450+3)}} = \sqrt{\frac{6 * 450 * 449}{448 * 451 * 453}} = 0,1151$$

$$SFK = 2 * 0,1151 * \sqrt{\frac{450^2 - 1}{(450-3)(450+5)}} = 2 * 0,1151 * \sqrt{\frac{202500 - 1}{447 * 455}} = 0,2297$$

10.3.2



Voortbouwend op het voorbeeld van de IQ-scores zoals dat hierboven is besproken. De scheefheid van de verdeling van die IQ-scores bedraagt 0,173 en de kurtosis 3,651. De standaardfout voor beide kengetallen hebben we hierboven reeds uitgerekend.

- Reconstrueer het 95% betrouwbaarheidsinterval voor de scheefheid. Wat kan je daaruit concluderen?
- Reconstrueer het 95% betrouwbaarheidsinterval voor de kurtosis. Wat kan je daaruit concluderen?

10.3.3



Het bestand "OLP functies.R" bevat een aantal functies waardoor we de betrouwbaarheidsintervallen voor beide kengetallen kunnen laten berekenen. De eerste twee functies die we hiervoor ontwikkelden zijn: `st.fout.Skew()` en `st.fout.Kurt()`. Deze twee functies stellen je in staat om respectievelijk de standaardfout van de scheefheid en van de kurtosis van de verdeling van een variabele te berekenen. In deze functies moet je verwijzen naar een variabele uit een dataset. Zo geeft het onderstaande fictieve voorbeeld je de standaardfout voor de scheefheid van de variabele V1 uit de dataset Data1:

```
> st.fout.Skew(Data1$V1)
```

Naast deze twee functie hebben we in dat bestand eveneens twee functies opgenomen die rechtstreeks betrouwbaarheidsintervallen als output geven: `betr.interval.Skew()` en `betr.interval.Kurt()`. Bij deze functies dien je te verwijzen naar een variabele uit je dataset. Standaard geven beide functies 95% betrouwbaarheidsintervallen. Met het argument `conf.level=` kan je opnieuw betrouwbaarheidsintervallen van een andere breedte opvragen. Hieronder twee fictieve toepassingen:

```
# Berekenen van een 95% betrouwbaarheidsinterval voor de
# scheefheid van een variabele V1 uit de fictieve dataset Data1
> betr.interval.Skew(Data1$V1)

# Berekenen van een 80% betrouwbaarheidsinterval voor een
# kurtosis van een variabele V1 uit de fictieve dataset Data1
> betr.interval.Kurt(Data1$V1, conf.level=0.80)
```

10.3.4 We hernemen het databestand Pirls2.Rdata.



- a) Wat zijn de 95% betrouwbaarheidsintervallen voor de scheefheid van de variabelen Informscore en Literatuurscore? En, wat kan daaruit concluderen?
- b) Wat zijn de 95% betrouwbaarheidsintervallen voor de kurtosis van de variabelen Informscore en Literatuurscore? En, wat kan daaruit concluderen?

10.4. Betrouwbaarheidsintervallen voor relatieve frequenties

i In dit OLP begonnen we de toepassing van beschrijvende statistiek met te stellen dat een eerste vorm van samenvatten simpelweg het tellen is van het aantal keren dat een bepaald kenmerk voorkomt: frequenties. Daarbij maakten we het onderscheid tussen absolute (aantal keer dat een kenmerk voorkomt) en relatieve frequenties (procentueel aantal keer dat een kenmerk voorkomt).

Ook voor deze aantallen beroepen we ons doorgaans op een steekproef om een schatting te krijgen van het aantal keer een bepaald kenmerk voorkomt in de populatie. Een typisch voorbeeld zijn de peilingen aanstaande stemgedrag die verschillende mediabedrijven uitvoeren. Daarbij hoor je vaak de commentaar: "peilingen zijn maar peilingen en kunnen een vertekend beeld geven". Een van de bronnen van een vertekend beeld is dat bij een peiling gebruik gemaakt wordt van een steekproef. Indien uit zo'n peiling blijkt dat een welbepaalde partij 35% van de stemmen zou krijgen, dan gaat dat om 35% van de stemmen van de personen die in de steekproef zaten. Zouden ze in de gehele populatie van stemgerechtigden ook 35% halen? Dat is maar de vraag. Rond dit percentage zit ook een zekere foutenmarge die we gaan kunnen bepalen. Net als bij alle overige kengetallen kunnen we rond relatieve frequenties een betrouwbaarheidsinterval bepalen.

Om een betrouwbaarheidsinterval te bepalen rond een relatieve frequentie hebben we een schatting nodig van de standaardfout. Zonder in te willen gaan op de statistisch/wiskundige achtergrond kunnen we stellen dat de standaardfout voor relatieve frequenties (of ook proporties) best als volgt geschat kan worden:

$$st.fout.proportie = \sqrt{\frac{p * (1 - p)}{n}}$$

Daarbij staat p gelijk aan de relatieve frequentie uitgedrukt als proportie (= een getal tussen nul en één) en staat n voor de steekproefomvang.

10.4.2 Het kan worden aangetoond dat, indien de steekproef voldoende groot is (lees vooral niet extreem klein is), de steekproevenverdeling van een relatieve frequentie de normaalverdeling volgt.



Nu je dit weet, bereken een betrouwbaarheidsinterval voor de twee proporties in de onderstaande tabel:

Tabel 10.1: Relatief aantal jongens en meisjes dat van zichzelf verklaard thuis nooit een boek te lezen

	%	n
Jongens	9,6%	294
Meisjes	3,2%	311

10.4.3



We hebben in het bestand "OLP functies.R" een functie opgenomen die de berekening van betrouwbaarheidsintervallen rond relatieve aantallen vergemakkelijkt: `betr.interval.Prop (p, n)`. Tussen de haakjes dien je te verwijzen naar het steekproefpercentage (uitgedrukt als proportie) (p) en de steekproefomvang (n). Bovendien kan je het extra argument `conf.level=` hanteren om een ander dan een 95% betrouwbaarheidsinterval te berekenen.

10.4.4 Herneem de gegevens in tabel 4.1.



- a) Bereken voor zowel jongens als meisjes opnieuw het 95%-betrouwbaarheids-interval, maar a.d.h.v. de specifieke functie in R.
- b) Bereken een 80%-betrouwbaarheidsinterval voor jongens en meisjes en verwoord de resultaten in eigen woorden.